Journal of Machine Learning Research 20 (2019) 1-28 Submitted 8/18; Revised 12/18; Published 1/19

# Iterated Learning in Dynamic Social Networks

**Bernard Chazelle** CHAZELLE@CS.PRINCETON.EDU
*Department of Computer Science, Princeton University,*
*35 Olden Street, Princeton, NJ 08544*

**Chu Wang*** CHUWANG@AMAZON.COM
*Amazon Inc.*
*500 Boren Avenue, Seattle, WA 98109*

**Editor:** Mehryar Mohri

## Abstract

A classic finding by (Kalish et al., 2007) shows that no language can be learned iteratively by rational agents in a self-sustained manner. In other words, if $A$ teaches a foreign language to $B$, who then teaches what she learned to $C$, and so on, the language will quickly get lost and agents will wind up teaching their own common native language. If so, how can linguistic novelty ever be sustained? We address this apparent paradox by considering the case of iterated learning in a social network: we show that by varying the lengths of the learning sessions over time or by keeping the networks dynamic, it is possible for iterated learning to endure forever with arbitrarily small loss.

## 1. Introduction

People typically form opinions by updating their current beliefs and reasons in response to new signals from other sources (friends, colleagues, social media, newspapers, etc.) (Tahbaz-Salehi et al., 2009; Acemoglu and Ozdaglar, 2011; Golub and Jackson, 2010). Suppose there were an information source that made a noisy version of the "truth" available to agents connected through a social network. Under which conditions would the agents reach consensus about their beliefs? What would ensure truthful consensus (meaning that the consensus coincided with the truth)? How long would it take for the process to converge? Addressing these questions requires agreeing on a formal model of distributed learning. Fully rational agents update their beliefs by assuming a prior and using Bayes' rule to integrate all past information available to them (Acemoglu et al., 2011; Mueller-Frank, 2013; Lobel and Sadler, 2015; Mossel et al., 2011; Banerjee, 1992; Bala and Goyal, 1998). Full rationality is intractable in practice (Molavi et al., 2015; Rahimian and Jadbabaie, 2016a), so much effort has been devoted to developing computationally effective mechanisms, including non- (or partially) Bayesian methods (Jadbabaie et al., 2012; Molavi et al., 2015; Golub and Jackson, 2010, 2012; Jadbabaie et al., 2013). Much of this line of work can be traced back to the seminal work of (DeGroot, 1974) on linear opinion pooling.

As the simplest example of iterated learning in a social network, consider a system consisting of one teacher and one learner. The teacher samples data from a distribution and sends it to the learner; the learner updates her belief via Bayes rule repeatedly in order to learn that distribution. This system is equivalent to the usual Bayesian inference scenario. Under mild assumptions, the learner will eventually learn the ground truth asymptotically (Gelman et al., 2013).

∗. Chu Wang did this work prior to joining Amazon

When the network structure comes into play, the dynamics of the learning process becomes more complicated. If the social network forms a chain $X, Y, Z, \ldots$ such that agents teach each other in sequence: X teaches Y, who then teaches Z, and so on, by a classic result of Griffiths and Kalish (Griffiths and Kalish, 2005), the information from the source will vanish after a finite number of iterations. At that point, the agents, assumed to be rational, will be "teaching" each other nothing they don't already know: iterated learning is not self-sustaining (Beppu and Griffiths, 2009; Griffiths and Kalish, 2007, 2005; Rafferty et al., 2009; Kirby et al., 2014; Griffiths et al., 2008; Perfors and Navarro, 2011; Rafferty et al., 2014; Smith, 2009; Kalish et al., 2007). Such findings are hard to validate empirically but variants of it are within the reach of experimental psychology (Kalish et al., 2007). Similar laboratory experiments with human subjects have confirmed the unstainability of iterated learning (Kalish et al., 2007; Beppu and Griffiths, 2009; Tamariz and Kirby, 2015; Bartlett; Griffiths et al., 2008). Similar results of unstainability are found in computational linguistics where, instead of agents sending information, the scenario is about language evolution through generations (Rafferty et al., 2009).

If agents interact in a more complicated and dynamic network, more possibilities of the learning dynamics will emerge. Dynamic networks are common occurrences in opinion dynamics (Hegselmann and Krause, 2002; Mohajer and Touri, 2013; Chazelle and Wang, 2016; Chazelle, 2015), but, to our knowledge, somewhat new in the context of social learning. Following the *Bayesian-Without-Recall* (BWR) model proposed by Rahimian and Jadbabaie in (Rahimian and Jadbabaie, 2016a), we assume the agents to be memoryless and rational: this means that they use Bayesian updates based on current beliefs and signals with no other information from the past, see also (Rahimian et al., 2015a,b; Rahimian and Jadbabaie, 2016b). The BWR model seeks to capture the benefits of rational behavior while keeping both the computation and the information stored to a minimum (Rahimian and Jadbabaie, 2016b).

We focus in this paper on sustainable learning: the conditions to ensure arbitrarily small information loss in truthful consensus (formal definition in the following sections). For chained learning, we show how keeping the length of the training sessions (number of samples transferred) growing slightly allows iterated learning to be sustained in perpetuity. This resolves the paradox raised from language evolution models (Kalish et al., 2007). We further analyze the case when the learner has the ability to reach back to her early ancestors for "fresher" data instead of listening to her direct ancestor, and show how this "hopped" learning mechanism further helps prevent information decay. For the case when the learning network changes over time, we show that under the assumption that each agent hears a noisy signal from the truth at a frequency bounded away from zero, the system reaches truthful consensus almost surely with a convergence rate polynomial in expectation. The relation between the convergence rate and the graph structure is also revealed with a seemingly counter-intuitive finding that agents in a better connected network learn more slowly.

We first introduce the iterated learning model framework with the notation, definitions, and basic properties in Section 2. Then we exam the scenarios of chained learning, hopped learning, and networked learning in Section 3, Section 4, and Section 5, respectively. We show our main results in each section, followed by details of the proofs and discussions.

## 2. Models, Preliminaries, and Notation

In this section, we formally define the problem mathematically. Assume there are $n$ agents denoted by their indices $1, 2, \ldots, n$. At time $t = 0, 1, \ldots$, the belief of agent $i$ is a probability distribution

over a state space $\Theta$, which is denoted by $\mu_{t,i}$. The interactions between agents are modeled by an infinite sequence $(G_t)_{t\geq 0}$, where each $G_t$ is a directed graph over the node set $\{1,\ldots,n\}$. An edge pointing from $i$ to $j$ in $G_t$ indicates that $i$ receives data from $j$ at time $t$. Intuitively, the direction of the edge has the same meaning as the "listen-to" activity. Typically, the sequence of graphs is specified ahead of time or is chosen randomly: the only condition that matters is that it should be independent of the randomness used in the data generating and learning process; specifically, taking expectations and variances of the random variables that govern the dynamics will assume a fixed graph sequence (possibly random). The adjacency matrix of $G_t$ is denoted by $A_t$: it is an $n \times n$ $0/1$ matrix.

### 2.1. The existence of an information source

When an information source exists whose belief is fixed, we label it agent $0$ and refer to it as the *truth*. In such a case, the graph $G_t$ is over the node set $\{0,1,\ldots,n\}$. Because agent $0$ (if it exists) holds the truth, no edge out of it points to another node. The adjacency matrix then becomes an $(n+1)\times(n+1)$ matrix whose first row is $(1,0,\ldots,0)$, with a self-loop at agent 0 for simplicity.

### 2.2. Data generation

At time $t \geq 0$, each agent $i > 0$ samples a state $\theta_{t,i} \in \mathbb{R}$ consistent with her own belief: $\theta_{t,i} \sim \mu_{t,i}$. A noisy measurement $a_{t,i} = \theta_{t,i} + \varepsilon_{t,i}$ is then sent to each agent $j$ such that $(A_t)_{ji} = 1$. All the noise terms $\varepsilon_{t,i}$ are sampled *iid* from $\mathcal{N}(0,\sigma^2)$. An equivalent formulation is to say that the likelihood function $l(a|\theta)$ is drawn from $\mathcal{N}(\theta,\sigma^2)$. In our setting, agent $i$ sends the same data to all of her neighbors; this is done for notational convenience and the same results would still hold if we were to resample independently for each neighbor. Except for the omission of explicit utilities and actions, our setting is easily identified as a variant of the BWR model of (Rahimian and Jadbabaie, 2016a).

### 2.3. Beliefs update

A single-step update for agent $i > 0$ consists of setting $\mu_{t+1,i}$ as the posterior $\mathbb{P}[\mu_{t+1,i}|d] \propto \mathbb{P}[d|\mu_{t,i}]\mathbb{P}[\mu_{t,i}]$, where $d$ is the data from the neighbors of $i$ received at time $t$. For the case when the beliefs are Gaussian, we get the classical update rules from Bayesian inference by plugging in the corresponding normal distribution (Box and Tiao, 2011). Updated beliefs remain Gaussian so we can use the notation $\mu_{t,i} \sim \mathcal{N}(x_{t,i}, \tau_{t,i}^{-1})$, where $\tau_{t,i}$ denotes the precision (inverse variance) $\sigma_{t,i}^{-2}$. Writing $\tau = \sigma^{-2}$ and letting $d_{t,i}$ denote the outdegree of $i$ in $G_t$, for any $i > 0$ and $t \geq 0$, we have

$$\begin{cases} x_{t+1,i} = (\tau_{t,i}x_{t,i} + \tau a_{t,1} + \cdots + \tau a_{t,d_{t,i}})/(\tau_{t,i} + d_{t,i}\tau); \\ \tau_{t+1,i} = \tau_{t,i} + d_{t,i}\tau, \end{cases} \tag{1}$$

where $a_{t,1},\ldots,a_{t,d_{t,i}}$ are the signals received by agent $i$ from its neighbors at time $t$.

### 2.4. The influence of the graph sequence

The graph sequence $G_t$ plays a crucial role in the dynamics of the system. A simple starting example is the constant graph of two agents with one directed edge. Such a graph sequence defines the case where one learner repeatedly gets samples from the information source. If the information source is regarded as the ground truth, this system is equivalent to the usual Bayesian inference

scenario. Instead of exhausting all possible graph sequences, in this paper, we focus on three representative types, namely the *chained learning*, *hopped learning*, and *networked learning* models. The fundamental problem we would like to solve is whether the system is able to converge to the *truth*; in other words, whether the truthful information is able to propagate uncorruptedly across the entire system.

## 3. Chained Learning

Following (Beppu and Griffiths, 2009; Griffiths and Kalish, 2007, 2005; Rafferty et al., 2009; Kirby et al., 2014; Griffiths et al., 2008; Perfors and Navarro, 2011; Rafferty et al., 2014; Smith, 2009; Kalish et al., 2007), we begin with *chained iterated learning*: a learner's state of belief is modeled by a distribution over a hypothesis space $\mathcal{H}$, which is itself equipped with a likelihood function: $\mathbb{P}[d|h]$ indicates the probability of generating data $d \in \mathcal{D}$ given the hypothesis $h \in \mathcal{H}$. A learner's state of belief may change after a learning process, and we naturally call her belief before and after the learning prior and posterior. Notice that the hypothesis space $\mathcal{H}$ is the same as the state space $\Theta$, but we will use $\mathcal{H}$ to emphasize the application background and avoid confusion. The initial hypothesis $\boldsymbol{h}_{\texttt{init}}$ generates $m_1$ items *iid* for the first learner. These items provide the training data $\mathbf{d}_1 = (d_{1,1}, \ldots, d_{1,m_1})$ with which the first learner Bayes-updates its prior. Its posterior is given by setting $t = 1$ in this formula:

$$\mathbb{P}[h|\mathbf{d}_t] = \mathbb{P}[\mathbf{d}_t|h]\,\mathbb{P}[h]/\mathbb{P}[\mathbf{d}_t], \quad \text{with } \mathbb{P}[\mathbf{d}_t] = \sum_{h\in\mathcal{H}} \mathbb{P}[\mathbf{d}_t|h]\,\mathbb{P}[h]. \tag{2}$$

From that point on, each successive learner updates its prior from their predecessor. For any $t > 1$, learner $t$ receives $m_t$ items sampled by the posterior of agent $t-1$ to form the training set $\mathbf{d}_t$. To do that, she picks a random hypothesis $h$ from $\mathcal{H}$ with probability $\mathbb{P}[h|\mathbf{d}_{t-1}]$ (the posterior of learner $t-1$) and then samples $m_t$ items *iid* from $h$ to form $\mathbf{d}_t \in \mathcal{D}^{m_t}$. The posterior $\mathbb{P}[h|\mathbf{d}_t]$ is derived according to (2). Note that learner $t$ has no direct access to the posterior of learner $t-1$ but only to data drawn from a hypothesis sampled from the posterior. Our formulation assumes a discrete space $\mathcal{H}$ but extends to continuous settings, as we show in §3.5.

In the case of linguistic transmission, each hypothesis $h \in \mathcal{H}$ is a "knob" whose setting is given by a number between 0 and 1, specifically the prior probability $\mathbb{P}[h]$. All learners share the same prior. Picking some $h$ from that prior specifies a *language* (also denoted $h$ for convenience). In this case, a language is defined as a probability distribution over $\mathcal{D}$, interpreted here as a set of *sentences*. In this way, the prior can be viewed as a mixture over $\mathcal{H}$: by abuse of terminology, we call it a *mixed* hypothesis, which we distinguish from a *pure* hypothesis of the form $h \in \mathcal{H}$ (corresponding to a single-point distribution). Access to language $h$ is achieved by random sampling: the sentence $d \in \mathcal{D}$ is picked with probability $\mathbb{P}[d|h]$.

Iterated learning proceeds as follows. After selecting language $h$ with probability $\mathbb{P}[h|\mathbf{d}_{t-1}]$, learner $t$ collects $m_t$ independent samples from $h$. Thus, given a tuple $\mathbf{d}_t = (d_1, \ldots, d_{m_t})$ of sentences from $\mathcal{D}$, the likelihood $\mathbb{P}[\mathbf{d}_t|h]$ is equal to $\prod_{1\le k\le m_t} \mathbb{P}[d_k|h]$. The learner is now ready to Bayes-update its prior. Of course, the first one $(t = 1)$ samples directly from the language $\boldsymbol{h}_{\texttt{init}}$ chosen for iterated learning. The notation is boldfaced to indicate that $\boldsymbol{h}_{\texttt{init}}$ may be a mixed hypothesis or, in other words, a distribution over hypotheses.

Suppose that $\mathcal{D} = \{d_1, \ldots, d_s\}$ and $\mathcal{H} = \{h_1, \ldots, h_n\}$ are both finite. After observing the data generated by the posterior of learner $t-1$, if learner $t$ winds up choosing $h_i$ then, by Bayesian updating, the probability $P_{ij}^{|t}$ that its posterior picks $h_j$ is given by:

$$P_{ij}^{|t} = \sum_{\mathbf{d} \in \mathcal{D}^{m_t}} \mathbb{P}[h_j|\mathbf{d}]\, \mathbb{P}[\mathbf{d}|h_i] = \sum_{\mathbf{d} \in \mathcal{D}^{m_t}} \frac{\mathbb{P}[\mathbf{d}|h_i]\, \mathbb{P}[\mathbf{d}|h_j]\, \mathbb{P}[h_j]}{\sum_{k=1}^{n} \mathbb{P}[\mathbf{d}|h_k]\, \mathbb{P}[h_k]}\,. \tag{3}$$

To our knowledge, the entire literature on the topic assumes a common, fixed sample size for all the learners: $m_t = m$. Equation (3) can be then interpreted as marginalizing a Gibbs sampler over the data space, which creates a Markov chain over the hypothesis space $\mathcal{H}$: if $\boldsymbol{h}^t$ denotes the row vector formed by the $n$ probabilities $\mathbb{P}[h_k \,|\, \mathbf{d}_t]$, then $\boldsymbol{h}^t = \boldsymbol{h}^{t-1} P^t$, where $\boldsymbol{h}^0 = \boldsymbol{h}_{\texttt{init}}$. Assuming ergodicity (in this case, a fairly inconsequential technical assumption), the chain can be shown to converge to a unique stationary distribution $\boldsymbol{h}$. It can be easily checked that it coincides with the prior: $\boldsymbol{h} = (\mathbb{P}[h_1], \ldots, \mathbb{P}[h_n])$ (Griffiths and Kalish, 2005; Norris, 1998); see (Rafferty et al., 2009, 2014) for an analysis of the mixing time in specific linguistic scenarios. This convergence reveals the long-term unsustainability of iterated learning. We show how diversifying the sample sizes $m_t$, hence making the Markov chain time-inhomogeneous, can overcome this weakness. In particular, we prove that it is sufficient for $m_t$ to increase logarithmically with respect to $t$ in order to achieve sustainability.

### 3.1. Self-Sustainability

We show how to make iterated learning self-sustaining in the presence of a finite hypothesis space $\mathcal{H} = \{h_1, \ldots, h_n\}$. This involves specifying a sequence of training session lengths $m_1, m_2, \ldots$ so that the posterior of any learner ends up differing from $\boldsymbol{h}_{\texttt{init}}$ by an arbitrarily small amount. Formally, given any $\delta, \varepsilon \geq 0$, we say that iterated learning is $(\delta, \varepsilon)$-*self-sustaining* if, with probability at least $1 - \varepsilon$, a random $h \in \mathcal{H}$ picked from any learner's posterior distribution differs from $\boldsymbol{h}_{\texttt{init}}$ in total variation by at most $\delta$. We recall a few facts: the hypothesis $h$ denotes a language modeled as a probability distribution over $\mathcal{D}$; the total variation distance is half the $\ell_1$-norm; and the posterior of learner $t$ after the $t$-th iteration is defined by marginalizing $\mathbb{P}[h|\mathbf{d}_t]$ over all samples $\mathbf{d}_t$ drawn from a random $h$ picked from the posterior of learner $t-1$ (or $\boldsymbol{h}_{\texttt{init}}$ if $t = 1$). As a shorthand, we speak of $\varepsilon$-self-sustainability to refer to the case $\delta = 0$.

The parameters $\delta$ and $\varepsilon$ allow us to distinguish between two metrics: the distance between two languages over $\mathcal{D}$ and the distance between two mixtures over $\mathcal{H}$. The two notions could differ widely. For example, if all of $\mathcal{H}$ corresponds to languages very close to $\boldsymbol{h}_{\texttt{init}}$, to achieve $(\delta, \varepsilon)$-self-sustainability might be easy for a tiny $\delta > 0$ but hopelessly difficult for $\delta = 0$. The complexity of iterated learning depends on the geometry of the languages formed by the pure hypotheses. This is best captured by introducing a metric that, though more specialized than the total variation (it works only on the simplex of probability vectors), brings all sorts of technical benefits: the *root-sine distance* between two probability distributions $\boldsymbol{u} = (u_1, \ldots, u_s)$ and $\boldsymbol{v} = (v_1, \ldots, v_s)$ over $\mathcal{D}$ is defined as

$$d_{RS}(\boldsymbol{u}, \boldsymbol{v}) = \sqrt{\frac{1}{2} \sum_{i,j=1}^{s} \left(\sqrt{u_i v_j} - \sqrt{u_j v_i}\right)^2} = \sqrt{1 - \left(\sum_{i=1}^{s} \sqrt{u_i v_i}\right)^2}\,. \tag{4}$$

Note that the *root-sine distance* will be used to measure similarities between two likelihoods, and we will continue the analysis of sustainability defined based on the total variation distance.

It would be surprising if this distance had not been used before, but we could not find a reference. We prove that it is indeed a metric in the Appendix and also explain its name. We show that it is related to the Hellinger, Bhattacharyya and total variation distances, $d_H$, $d_B$, $d_{TV}$ by the following relations:

$$\begin{cases} d_H = \sqrt{1 - \sqrt{1 - d_{RS}^2}}\,; \\ d_B = -\frac{1}{2}\ln(1 - d_{RS}^2)\,; \\ d_{TV} \le \sqrt{2s}\, d_{RS}. \end{cases} \tag{5}$$

### 3.2. The results

We focus on the "pure" case $\boldsymbol{h}_{\texttt{init}} \in \mathcal{H}$, and later briefly discuss how to generalize the method to mixed hypotheses. Using the shorthand $\mathtt{d}_{ij}$ for $d_{RS}(\mathbb{P}[\cdot|h_i], \mathbb{P}[\cdot|h_j])$, we define $\mathtt{d}_i := \min_{j:j\neq i} \mathtt{d}_{ij}$. Let $\boldsymbol{p} = (p_1, p_2, \ldots, p_n)$ be the prior distribution over $\mathcal{H}$, where $p_i := \mathbb{P}[h_i]$. We can obviously assume that each $p_i$ is positive and that all the pure hypotheses are distinct, hence $\mathtt{d}_i > 0$. The two theorems below assume that $\boldsymbol{h}_{\texttt{init}} = h_1$.

**Theorem 1**. *For any positive $\varepsilon < 1$, the following sample size sequence makes iterated learning $\varepsilon$-self-sustaining:*

$$m_t = \frac{4}{\mathtt{d}_1^2} \ln \frac{nt}{\varepsilon\, p_1} = \frac{4}{\mathtt{d}_1^2}\Big(\log \frac{t}{\varepsilon} + C\Big),$$

*for some $C > 0$ independent of $t, \varepsilon, \mathtt{d}_1$.*

The factor $4$ can be reduced to $2^{1+o(1)}$ if we adjust the constant $C$. It is to be expected that the lengths of the training sessions should grow to infinity as $p_1$ tends to zero, as the vanishing prior makes it increasingly difficult for the posteriors to "attach" to $h_1$. The session lengths are sensitive to the minimum distance between the languages specified by $\mathcal{H}$ and the target language $h_1$. Settling for $(\delta, \varepsilon)$-self-sustainability allows us to remove this dependency.

**Theorem 2**. *For any positive $\delta, \varepsilon < 1$, the following sample size sequence makes iterated learning $(\delta, \varepsilon)$-self-sustaining:*

$$m_t = \frac{8sn^2}{\delta^2}\Big(\ln \frac{t}{\varepsilon} + C\Big).$$

*for some $C > 0$ independent of $t, \delta, \varepsilon$.*

### 3.3. The proofs

To establish Theorem 1, we estimate the probability $P^*$ that each leaner ends up picking $h_1$. Recall that $\boldsymbol{h}^t$ is the posterior distribution of learner $t$, by the Markovian property of the system,

$$P^* = \mathbb{P}[\boldsymbol{h}^0 = h_1] \prod_{t\geq 0} \mathbb{P}[\boldsymbol{h}^{t+1} = h_1 | \boldsymbol{h}^t = h_1] = \prod_{t\geq 1} P_{11}^{|t}. \tag{6}$$

Since the matrix $P^{|t}$ is the transition matrix of a Markov chain, we proceed by bounding its off-diagonal elements $P_{ij}^{|t}$ for $i \neq j$. We have

$$\begin{aligned} P_{ij}^{|t} &\leq \sum_{\mathbf{d}\in\mathcal{D}^{m_t}} \frac{\mathbb{P}[\mathbf{d}|h_i]\,\mathbb{P}[\mathbf{d}|h_j]p_j}{\mathbb{P}[\mathbf{d}|h_i]p_i + \mathbb{P}[\mathbf{d}|h_j]p_j} = \frac{p_j}{p_i} \sum_{\mathbf{d}\in\mathcal{D}^{m_t}} \frac{\big(\frac{p_i}{p_j}\big)\mathbb{P}[\mathbf{d}|h_i]\,\mathbb{P}[\mathbf{d}|h_j]}{\big(\frac{p_i}{p_j}\big)\mathbb{P}[\mathbf{d}|h_i] + \mathbb{P}[\mathbf{d}|h_j]} \\ &\leq \frac{1}{2}\sqrt{\frac{p_j}{p_i}} \sum_{\mathbf{d}\in\mathcal{D}^{m_t}} \sqrt{\mathbb{P}[\mathbf{d}|h_i]\,\mathbb{P}[\mathbf{d}|h_j]} = \frac{1}{2}\sqrt{\frac{p_j}{p_i}} \Big(\sum_{d\in\mathcal{D}} \sqrt{\mathbb{P}[d|h_i]\,\mathbb{P}[d|h_j]}\Big)^{m_t} \\ &\leq \frac{1}{2}\sqrt{\frac{p_j}{p_i}} \exp\left\{\frac{m_t}{2}\Big(\Big(\sum_{d\in\mathcal{D}} \sqrt{\mathbb{P}[d|h_i]\,\mathbb{P}[d|h_j]}\Big)^2 - 1\Big)\right\}, \end{aligned}$$

where the two equalities are simple deformations, the first inequality is achieved by dropping some non-negative terms from the definition of $P_{ij}^{|t}$ in (3), the second inequality is obtained via Young's inequality, and the last inequality is from Taylor expansion of the natural logarithm function at 1. By definition of the root-sine distance, we have

$$P_{ij}^{|t} \leq \frac{1}{2}\sqrt{\frac{p_j}{p_i}} e^{-\frac{1}{2}\mathtt{d}_{ij}^2 m_t} \quad (i \neq j). \tag{7}$$

Setting $i = 1$ in (7) and summing over $2 \leq j \leq n$, it follows by Cauchy-Schwarz that

$$\sum_{j=2}^{n} P_{1j}^{|t} \leq \frac{1}{2}\sqrt{\frac{n(1-p_1)}{p_1}}\, e^{-\frac{1}{2}\mathtt{d}_1^2 m_t}. \tag{8}$$

Combining (6) and (8) yields

$$P^* \geq \prod_{t\geq 1}\left(1 - \frac{1}{2}\sqrt{\frac{n(1-p_1)}{p_1}}\, e^{-\frac{1}{2}\mathtt{d}_1^2 m_t}\right) \geq 1 - \frac{1}{2}\sqrt{\frac{n(1-p_1)}{p_1}} \sum_{t\geq 1} e^{-\frac{1}{2}\mathtt{d}_1^2 m_t}. \tag{9}$$

Given $0 < \varepsilon < 1$, we constrain the sequence $(m_t)$ to satisfy:

$$\sum_{t\geq 1} e^{-\frac{1}{2}\mathtt{d}_1^2 m_t} < \varepsilon\sqrt{\frac{4p_1}{n(1-p_1)}}. \tag{10}$$

For example, we can pick the sequence

$$m_t = \frac{1}{\mathtt{d}_1^2} \ln \frac{n(1-p_1)t^4}{\varepsilon^2 p_1},$$

which completes the proof. A closer look at the calculation shows that the factor $t^4$ can be reduced to $C_\alpha t^{2+\alpha}$ for any small $\alpha > 0$ and a suitable constant $C_\alpha > 0$, which makes the dependency on $t$ arbitrarily close to $(2/\mathtt{d}_1^2) \ln t$.

To prove Theorem 2, we set a target distance $\rho := \delta/(n\sqrt{2s})$ and find a subset $A \subseteq \mathcal{H}$ such that (i) $\mathtt{d}_{1j} \le \rho n$ for $j \in A$ and (ii) $\mathtt{d}_{ij} \ge \rho$ for $i \in A$ and $j \notin A$. To see why such a subset must exist, consider spheres centered at $\boldsymbol{h}_{\mathtt{init}} = h_1$ of radius $k\rho$, for $k = 1, \ldots, n+1$ (with respect to $d_{RS}$). These define $n+1$ disjoint (open) regions and, by the pigeonhole principle, at least one of them must be empty. We set $A$ to include all the points in the regions preceding the empty one; note that $h_1 \in A$. The claim follows from the triangular inequality. We begin with a straightforward generalization of (8): for any $i \in A$,

$$\sum_{j \notin A} P_{ij}^{|t} \le \frac{1}{2}\sqrt{\frac{n(1-p_A)}{p_A}}\, e^{-\frac{1}{2}\rho^2 m_t}, \tag{11}$$

where $p_A := \min_{i \in A} p_i$. Now let $P^*$ be the probability that $\boldsymbol{h}^t \in A$ for each $t$, then (6) and (9) are generalized to

$$P^* \ge \prod_{t \ge 1} \left(1 - \max_{i \in A} \sum_{j \notin A} P_{ij}^{|t}\right) \ge 1 - \frac{1}{2}\sqrt{\frac{n(1-p_A)}{p_A}} \sum_{t \ge 1} e^{-\frac{1}{2}\rho^2 m_t}. \tag{12}$$

Setting

$$m_t = \frac{1}{\rho^2} \ln \frac{n(1-p_A)t^4}{\varepsilon^2 p_A} \tag{13}$$

ensures that $P^* > 1 - \varepsilon$. The root-sine distance between the languages denoted by $h_1$ and any $h \in A$ is at most $\rho n$, so that, by (5), the total variation distance is bounded by $\sqrt{2s}\rho n = \delta$, which concludes the proof of Theorem 2.

So far, we have analyzed only the "pure" case $\boldsymbol{h}_{\mathtt{init}} \in \mathcal{H}$. The idea of the training is to prevent the prior to "drag" the posterior mixture all across $\mathcal{H}$. It should be clear that a similar result obtains if $\boldsymbol{h}_{\mathtt{init}} \in \Delta\mathcal{H}$ is concentrated on a subset $A$ of $\mathcal{H}$. The proof follows the path charted in Theorem 2 and need not be repeated here. It is crucial to note, however, that this result is to be understood in a coarse-graining sense: iterated learning cannot ensure that the original weights in the mixture $\boldsymbol{h}_{\mathtt{init}}$ are retained but only that $A$ contributes most of the mass in the posteriors. To retain the weights would require changing the stationary distribution to conform with $\boldsymbol{h}_{\mathtt{init}}$, as the process unfolds, something that straightforward Bayesian learning seems unable to do. Learning pure hypotheses bypasses that difficulty.

### 3.4. Applications

We briefly discuss a direct application of our results to a well-known model of language acquisition via iterated learning and we mention some natural extensions of the techniques.

**Language evolution.** Rafferty et al. show how iterated learning fails rapidly in a simple model of language evolution (Rafferty et al., 2009). Given $n$ hypotheses, iterated learning with fixed-length training sessions ceases to learn anything new after only $O(\log n \log\log n)$ rounds. Our previous theorems show how to turn this around and achieve self-sustainality. In our notational system, their model is defined on a hypothesis space $\mathcal{H} = \{h_1, \ldots, h_n\}$, where $n = 2^k$ and $h_i$ denotes the

language whose sentences are words in $\{0,1,?\}^k$ with exactly $m$ question marks and $0,1$ matching the binary decomposition of $i-1$ outside the question marks. For example, if $k=4$ and $m=2$, then $h_3$ denotes the language

$$\{\,00??, 0?1?, ?01?, 0??0, ?0?0, ??10\,\}.$$

We can assume that $m$ is much smaller than $k$. Each language has the same length $\binom{k}{m}$ and the total number of sentences is $s=\binom{k}{m}2^{k-m}$. The prior is given by $\mathbb{P}[h_i]=p_i=1/n$. Given a hypothesis $h_i$, $\mathbb{P}[d|h_i]=1/\binom{k}{m}$ if $d$ has $m$ question marks and match the bits of $i-1$ elsewhere; else it is 0 (and $d,h$ are called incompatible). Given $h\in\mathcal{H}$,

$$\begin{cases} \mathbb{P}[d]=\sum_{h\in\mathcal{H}}\mathbb{P}[d|h]\mathbb{P}[h]=2^{m-k}/\binom{k}{m}\,; \\ \mathbb{P}[h|d]=\mathbb{P}[d|h]\mathbb{P}[h]/\mathbb{P}[d]=2^{-m} \quad \text{(or 0 if } d,h \text{ are incompatible).} \end{cases}$$

We easily check that $\mathtt{d}_1^2=1-\big(\sum_{i=1}^s\sqrt{a_ib_i}\,\big)^2\geq 1-\big(\frac{m}{k}\big)^2>\frac{1}{2}$; hence, by Theorem 1, session lengths $m_t$ no larger than $O(\log\frac{t}{\varepsilon})$ are sufficient to maintain $\varepsilon$-self-sustainability.

**Meanings and utterances.** In the use of iterated learning for studying language evolution (Griffiths and Kalish, 2005; Perfors and Navarro, 2011), it is common to model the data $\mathbf{d}$ as a joint distribution $(\boldsymbol{x},\boldsymbol{y})$ over a product space $\mathcal{X}^{m_t}\times\mathcal{Y}^{m_t}$. The idea is to distinguish between "meanings" $\boldsymbol{x}$ and "utterances" $\boldsymbol{y}$. In this setting, $\mathbb{P}[\mathbf{d}|h]=\mathbb{P}[\boldsymbol{y}|\boldsymbol{x},h]\mu(\boldsymbol{x})$, where $\mu(\boldsymbol{x})$ is the probability of generating $\boldsymbol{x}$. The transition matrix of the Markov chain thus becomes

$$\begin{aligned} P_{ij}^{|t} &= \sum_{\boldsymbol{x}\in\mathcal{X}^{m_t}}\sum_{\boldsymbol{y}\in\mathcal{Y}^{m_t}}\mathbb{P}[h_j|\boldsymbol{x},\boldsymbol{y}]\,\mathbb{P}[\boldsymbol{y}|\boldsymbol{x},h_i]\mu(\boldsymbol{x}) \\ &= \sum_{\boldsymbol{x}\in\mathcal{X}^{m_t}}\sum_{\boldsymbol{y}\in\mathcal{Y}^{m_t}}\frac{\mathbb{P}[\boldsymbol{y}|\boldsymbol{x},h_i]\,\mathbb{P}[\boldsymbol{y}|\boldsymbol{x},h_j]\,\mathbb{P}[h_j]}{\sum_{k=1}^m\mathbb{P}[\boldsymbol{y}|\boldsymbol{x},h_k]\,\mathbb{P}[h_k]}\mu(\boldsymbol{x})\,. \end{aligned} \tag{14}$$

Since the output $\boldsymbol{y}$ now depends on both the hypothesis and the input data, we redefine $\mathtt{d}_{ij}$ as the root-sine distance between the two distributions $\mathbb{P}[\boldsymbol{y}|\boldsymbol{x},h_i]\mu(\boldsymbol{x})$ and $\mathbb{P}[\boldsymbol{y}|\boldsymbol{x},h_j]\mu(\boldsymbol{x})$:

$$\big(\mathtt{d}'_{ij}\big)^2 := 1-\left(\sum_{\boldsymbol{x}\in\mathcal{X}}\sum_{\boldsymbol{y}\in\mathcal{Y}}\sqrt{\mathbb{P}[\boldsymbol{y}|\boldsymbol{x},h_i]\mathbb{P}[\boldsymbol{y}|\boldsymbol{x},h_j]}\,\mu(\boldsymbol{x})\right)^2 \tag{15}$$

and we define $\mathtt{d}'_i:=\min_{j:j\neq i}\mathtt{d}'_{ij}$. Given any $i\neq j$,

$$\begin{aligned} P_{ij}^{|t} &\leq \sum_{\boldsymbol{x}\in\mathcal{X}^{m_t}}\sum_{\boldsymbol{y}\in\mathcal{Y}^{m_t}}\frac{\mathbb{P}[\boldsymbol{y}|\boldsymbol{x},h_i]\,\mathbb{P}[\boldsymbol{y}|\boldsymbol{x},h_j]\,p_j}{\mathbb{P}[\boldsymbol{y}|\boldsymbol{x},h_i]\,p_i+\mathbb{P}[\boldsymbol{y}|\boldsymbol{x},h_j]\,p_j}\mu(\boldsymbol{x}) \\ &\leq \frac{1}{2}\sqrt{\frac{p_j}{p_i}}\sum_{\boldsymbol{x}\in\mathcal{X}^{m_t}}\sum_{\boldsymbol{y}\in\mathcal{Y}^{m_t}}\sqrt{\mathbb{P}[\boldsymbol{y}|\boldsymbol{x},h_i]\,\mathbb{P}[\boldsymbol{y}|\boldsymbol{x},h_j]}\,\mu(\boldsymbol{x}) \\ &\leq \frac{1}{2}\sqrt{\frac{p_j}{p_i}}\Big(\sum_{\boldsymbol{x}\in\mathcal{X}}\sum_{\boldsymbol{y}\in\mathcal{Y}}\sqrt{\mathbb{P}[\boldsymbol{y}|h_i]\,\mathbb{P}[\boldsymbol{y}|h_j]}\,\mu(\boldsymbol{x})\Big)^{m_t} \\ &\leq \frac{1}{2}\sqrt{\frac{p_j}{p_i}}\exp\left\{\frac{m_t}{2}\Big(\Big(\sum_{\boldsymbol{x}\in\mathcal{X}}\sum_{\boldsymbol{y}\in\mathcal{Y}}\sqrt{\mathbb{P}[\boldsymbol{y}|\boldsymbol{x},h_i]\,\mathbb{P}[\boldsymbol{y}|\boldsymbol{x},h_j]}\,\mu(\boldsymbol{x})\Big)^2-1\Big)\right\}. \end{aligned}$$

This gives us this new version of inequality (7), which we can use as the basis for a repeat of the argument of the previous section:

$$P_{ij}^{|t} \le \frac{1}{2}\sqrt{\frac{p_j}{p_i}} e^{-\frac{1}{2}\mathtt{d}_{ij}'^2 m_t} \quad (i \ne j). \tag{16}$$

### 3.5. Gaussian chained learning

When iterated learning operates over a hypothesis space $\mathcal{H}$ parametrized continuously, say, in $\mathbb{R}$, the minimum root-sine distance usually vanishes and the previous arguments run into singularities and collapse. A new approach is needed. To make our discussion concrete, we assume that the prior distribution of each learner is a Gaussian $\mathbb{P}[h] \sim N(\bar{\mu}, \bar{\sigma}^2)$ and that the likelihood of producing data $d$ given hypothesis $h$ is also normal: $\mathbb{P}[d|h] = N(h, \sigma^2)$. The likelihood can also be understood as a noisy measurement of $h$: $d = h + \phi$, where the noise $\phi \sim N(0, \sigma^2)$. We assume that the data received by the first learner comes from $N(\mu_0, \sigma_0^2)$. This is the simplest instance of a continuous setting in which the root-sine distance argument fails. We discuss it in some detail, considering both chained learning and its generalizations; and then we use the results to treat the case of iterated Bayesian linear regression.

During its training session, the $t$-th learner receives data $\mathbf{d}_t = (d_{t,1}, \ldots, d_{t,m_t})$ from its predecessor: it is obtained by first picking a random hypothesis $h$ from the posterior of learner $t-1$ and then collecting $m_t$ independent random samples from $N(h, \sigma^2)$. For the case $t = 1$, we can treat the original teacher as learner 0 with its posterior equal to $N(\mu_0, \sigma_0^2)$. Learner $t$ Bayes-updates its posterior as follows:

$$\mathbb{P}[h|\mathbf{d}_t] \propto \mathbb{P}[\mathbf{d}_t|h]\mathbb{P}[h] \propto \exp\Big(-\frac{1}{2\sigma^2}\sum_{i=1}^{m_t}(d_{t,i}-h)^2\Big)\exp\Big(-\frac{1}{2\bar{\sigma}^2}(h-\bar{\mu})^2\Big),$$

which is still Gaussian, with mean and variance denoted by $\mu_t$ and $\sigma_t^2$, respectively. Carrying out the usual square completion gives up these update rules: for $t > 0$,

$$\begin{cases} \mu_t = \frac{1}{\bar{\tau}+m_t\tau}\left(\bar{\tau}\bar{\mu} + \tau(d_{t,1} + d_{t,2} + \cdots + d_{t,m_t})\right) \\ \tau_t \ = \bar{\tau} + m_t\tau, \end{cases} \tag{17}$$

where we define the precisions $\tau = 1/\sigma^2$, $\bar{\tau} = 1/\bar{\sigma}^2$, and $\tau_t = 1/\sigma_t^2$. We say that iterated learning is *$\varepsilon$-self-sustaining* if $|\mathbb{E}\,\mu_t - \mu_0| \le \varepsilon$ and $\sigma_t^2 + \operatorname{var}\mu_t$ remains bounded for all $t$. If $\sigma_t^2 + \operatorname{var}\mu_t \to 0$ as $t \to \infty$, we say that iterated learning is *strongly $\varepsilon$-self-sustaining*. We consider successively the case of chained iterated learning and the more challenging "hopping" scenario in which a new learner picks a random teacher from the past (instead of the previous one).

In chained iterated learning, the data $d_{t,i}$ is a noisy message drawn from the posterior of the $(t-1)$-th learner; hence $d_{t,i} \sim N(\mu_{t-1}, \sigma_{t-1}^2 + \sigma^2)$. In view of (17), $\mu_t$ is itself Gaussian. By taking the expectation and variance of equation (17), we find the following recursive relations for $\mathbb{E}\,\mu_t$ and $\operatorname{var}\mu_t$: for $t > 0$,

$$\begin{cases} \mathbb{E}\,\mu_t = \frac{1}{\bar{\tau}+m_t\tau}\big(\bar{\tau}\bar{\mu} + m_t\tau\,\mathbb{E}\,\mu_{t-1}\big); \\ \operatorname{var}\mu_t = \frac{m_t\tau^2}{(\bar{\tau}+m_t\tau)^2}\big(\operatorname{var}\ \mu_{t-1} + \sigma_{t-1}^2 + \sigma^2\big). \end{cases} \tag{18}$$

If we define $\beta_t := m_t\tau/(\bar{\tau} + m_t\tau)$, then (18) becomes $\mathbb{E}\,\mu_t = \beta_t\,\mathbb{E}\,\mu_{t-1} + (1-\beta_t)\bar{\mu}$. If $m_t = m$ is a constant, then so is $\beta_t$, and the recursive relation (18) becomes

$$\mathbb{E}\,\mu_t - \bar{\mu} = \beta_1^t(\mu_0 - \bar{\mu}),$$

which shows that $\mathbb{E}\,\mu_t$ converges to $\bar{\mu}$ exponentially fast. As in the discrete case, iterated learning is not self-sustainable with constant-length training sessions. By letting $m_t$ increase as $t^{1+o(1)}$ order, however, we can achieve self-sustainability:

**Theorem 3** *For any $0 < \varepsilon < 1$, the following sample size sequence makes chained iterated learning strongly $\varepsilon$-self-sustaining:*

$$m_t = \frac{|\mu_0 - \bar{\mu}|}{\varepsilon}\Big(1 + \frac{1}{c}\Big)\Big(\frac{\sigma}{\bar{\sigma}}\Big)^2 t^{1+c},$$

*for an arbitrarily small constant $c > 0$.*

**Proof** We observe that $\mathbb{E}\,\mu_t$ is a convex combination of $\bar{\mu}$ and $\mathbb{E}\,\mu_s$ $(s < t)$; specifically,

$$\mathbb{E}\,\mu_t = \prod_{s=1}^{t}\beta_s\mu_0 + \Big(1 - \prod_{s=1}^{t}\beta_s\Big)\bar{\mu}. \tag{19}$$

Because $\sum_{s>0}(1/s)^{1+c} < 1 + \int_1^\infty x^{-1-c}\,dx = 1 + 1/c$, we have

$$\begin{aligned} 1 \;&\geq\; \prod_{s=1}^{t}\beta_s = \prod_{s=1}^{t}\Big(1 - \frac{\bar{\tau}}{m_s\tau + \bar{\tau}}\Big) \geq 1 - \sum_{s=1}^{t}\frac{\bar{\tau}}{m_s\tau + \bar{\tau}} \\ &\geq\; 1 - \frac{\varepsilon}{|\mu_0 - \bar{\mu}|}\Big(\frac{c}{c+1}\Big)\sum_{s=1}^{\infty}\frac{1}{s^{1+c}} > 1 - \frac{\varepsilon}{|\mu_0 - \bar{\mu}|}\,. \end{aligned}$$

This shows that

$$|\mathbb{E}\,\mu_t - \mu_0| = \Big(1 - \prod_{s=1}^{t}\beta_s\Big)|\bar{\mu} - \mu_0| \leq \varepsilon.$$

By (17), $\sigma_t^2 = 1/\tau_t < 1/m_t\tau \to 0$. Since $\sigma_{t-1}^2 \leq \bar{\sigma}^2$ for $t > 1$, it follows from (18) that $\operatorname{var}\mu_t \leq (\operatorname{var}\mu_{t-1} + \sigma^2 + \bar{\sigma}^2)/m_t$ for $t > 1$, and $\operatorname{var}\mu_1 \leq (\sigma_0^2 + \sigma^2)/m_1$. Writing $M_t := m_t m_{t-1}\ldots m_1$, we have

$$\begin{aligned} M_t\operatorname{var}\mu_t &\leq M_{t-1}\operatorname{var}\mu_{t-1} + M_{t-1}(\sigma^2 + \bar{\sigma}^2) \\ &\leq tM_{t-1}(\sigma_0^2 + \sigma^2 + \bar{\sigma}^2), \end{aligned}$$

and thus $\operatorname{var}\mu_t \leq (\sigma_0^2 + \sigma^2 + \bar{\sigma}^2)t/m_t \to 0$ since $m_t = \Omega(t^{1+c})$. ■

### 3.6. Iterated Bayesian Linear Regression

The iterated version of Bayesian linear regression has been the subject of extensive study in the field of psychology (Kalish et al., 2007; Beppu and Griffiths, 2009; Tamariz and Kirby, 2015; Bartlett; Griffiths et al., 2008). The work has involved experimentation with human subjects but little in

the way of theoretical analysis. This section is a first step toward filling this gap. The task at hand is to estimate a hypothesis $h \in \mathcal{H} := \mathbb{R}^d$ given a noisy measurements on the hyperplane $y = h^T x$, where $x \in \mathbb{R}^d$. In the Bayesian setting, we assume a Gaussian prior on the hypothesis space: $\mathbb{P}[h] \sim N(\bar{\mu}, \bar{\sigma}^2 I_d)$. The data is given by $(x, y)$, where $x \sim N(0, I_d)$ and $y = h^T x + \phi$, for $\phi \sim N(0, \sigma^2)$ (with $x, \phi$ independent). Since we typically make several measurements, we write this (likelihood) relation in matrix form: $y = Xh + \phi$, where $y \in \mathbb{R}^m$ (with $m$ the number of measurements); $\phi \sim N(0, \sigma^2 I_m)$; and $X$ is an $m$-by-$d$ matrix each of whose rows denotes a random vector $x \sim N(0, I_d)$. This means that the matrix $X$ is random (a fact of key importance in our discussion below). We have:

$$\begin{cases} \mathbb{P}[\phi] \sim \exp\{-\frac{1}{2\sigma^2}\|\phi\|_2^2\} & \texttt{(noise)} \\ \mathbb{P}[h] \sim \exp\{-\frac{1}{2\bar{\sigma}^2}\|h - \bar{\mu}\|_2^2\} & \texttt{(prior)} \\ \mathbb{P}[y|X,h] \sim \exp\{-\frac{1}{2\sigma^2}\|y - Xh\|_2^2\} & \texttt{(likelihood)} \end{cases}$$

In iterated Bayesian linear regression, the $t$-th learner receives her data from learner $t-1$. Here, learner 0 is treated just like any other agent, except that his prior $\mathbb{P}[h] \sim N(\mu_0, \bar{\sigma}^2 I_d)$ is the distribution to be learned iteratively. Since sampling from the prior is independent of $X$, Bayesian updating gives the posterior $N(\mu_t, \Sigma_t)$, where

$$\mathbb{P}[h|X,y] = \mathbb{P}[h]\,\mathbb{P}[y|X,h]/\mathbb{P}[y|X] \sim \exp\Big\{-\frac{1}{2\bar{\sigma}^2}\|h - \bar{\mu}\|_2^2 - \frac{1}{2\sigma^2}\|y - Xh\|_2^2\Big\}.$$

Completing the square in the usual fashion shows that the posterior of learner $t$ is given by:

$$\begin{cases} \Sigma_t = \left(\bar{\sigma}^{-2} I_d + \sigma^{-2} X_t^T X_t\right)^{-1}; \\ \mu_t = \Sigma_t \left(\bar{\sigma}^{-2}\bar{\mu} + \sigma^{-2} X_t^T y_t\right), \end{cases} \tag{20}$$

where $(X_t, y_t)$ is the data gathered by learner $t$ from her predecessor: specifically, $y_t = X_t h + \phi_t$, where $h$ is collected from the $(t-1)$-th learner by sampling his posterior distribution $N(\mu_{t-1}, \Sigma_{t-1})$.

**Theorem 4** *Given any small enough $\delta, \varepsilon > 0$, the following sample size sequence for iterated Bayesian linear regression ensures that $\|\mathbb{E}\mu_t - \mu_0\|_2 \le \delta$ with probability greater than $1 - \varepsilon$:*

$$m_t = D_c \frac{\|\mu_0 - \bar{\mu}\|_2}{\delta} \Big(\frac{\sigma}{\bar{\sigma}}\Big)^2 t^{1+c} + D_c\, d \log \frac{t+1}{\varepsilon},$$

*for an arbitrarily small $c > 0$ and a constant $D_c$ that depends only on $c$.*

**Proof** We proceed in two steps: first, we show that to keep $\mathbb{E}\mu_t$ arbitrarily close to $\mu_0$ for all $t$ hinges on spectral properties of certain random matrices; second, we call on known facts about the singular values of random Gaussian matrices to translate the spectral condition into a high-probability event. The proof unfolds as a series of simple relations, which we state first and then demonstrate. The first one follows directly from (20):

$$\mathbb{E}\mu_t = (I_d + M_t)^{-1} (\bar{\mu} + M_t\, \mathbb{E}\mu_{t-1}), \quad \text{where} \quad M_t := \Big(\frac{\bar{\sigma}}{\sigma}\Big)^2 X_t^T X_t. \tag{21}$$

Note that (21) is a randomized recursive relation since the data points $X_1, X_2, \ldots$ are themselves random. We note that all the matrices whose inverses are taken are positive definite, hence nonsingular. To move on to our second relation, we define the matrix

$$Q_t := (I_d + M_t)^{-1} M_t (I_d + M_{t-1})^{-1} M_{t-1} \cdots (I_d + M_1)^{-1} M_1,$$

for $t > 0$, with $Q_0 = I_d$, and prove by induction that

$$\mathbb{E}\,\mu_t = Q_t \mu_0 + (I_d - Q_t)\bar{\mu}. \tag{22}$$

The base case is obvious so we assume that $t > 0$: by (21),

$$\begin{aligned}
\mathbb{E}\,\mu_t &= (I_d + M_t)^{-1}(\bar{\mu} + M_t\,\mathbb{E}\,\mu_{t-1}) \\
&= (I_d + M_t)^{-1}(\bar{\mu} + M_t Q_{t-1}\mu_0 + M_t(I_d - Q_{t-1})\bar{\mu}) \\
&= (I_d + M_t)^{-1} M_t Q_{t-1}\mu_0 + (I_d + M_t)^{-1}(I_d + M_t(I_d - Q_{t-1}))\bar{\mu} \\
&= Q_t\,\mu_0 + (I_d - (I_d + M_t)^{-1} M_t Q_{t-1})\bar{\mu},
\end{aligned}$$

which proves (22). Our next goal is to bound the information decay $\|\mathbb{E}\,\mu_t - \mu_0\|_2$. To do that, we investigate the spectral norm of the matrix $I_d - Q_t$, which leads to our third relation. We prove by induction that, for $t > 0$,

$$\|I_d - Q_t\|_2 \le \sum_{s=1}^{t} \|A_s\|_2, \tag{23}$$

where $A_s := (I_d + M_s)^{-1}$. For $t = 1$, $Q_1 = (I_d + M_1)^{-1} M_1 = I_d - (I_d + M_1)^{-1}$ and the claim follows. If $t > 1$, then

$$\begin{aligned}
\|I_d - Q_t\|_2 &= \|(I_d - Q_{t-1}) + (Q_{t-1} - Q_t)\|_2 \\
&\le \|I_d - Q_{t-1}\|_2 + \|Q_t - Q_{t-1}\|_2 \le \sum_{s=1}^{t-1} \|A_s\|_2 + \|\Psi\|_2,
\end{aligned}$$

where $\Psi := (A_t M_t - I_d) Q_{t-1}$. Since $A_t(I_d + M_t) = I_d$, we have $\Psi = -A_t Q_{t-1}$. Each matrix $M_s$ is positive semidefinite, so the eigenvalues of $(I_d + M_s)^{-1} M_s$ are of the form $\lambda/(1+\lambda)$, where $\lambda \ge 0$. This shows that all the eigenvalues of $Q_s$ are between 0 and 1; therefore $\|Q_s\|_2 \le 1$. The eigenvalues of $I_d - A_t M_t$ are the same as those of $A_t$; hence, by submultiplicativity, $\|\Psi\|_2 \le \|A_t\|_2 \|Q_{t-1}\|_2 \le \|A_t\|_2$, which establishes (23).

We are now ready to express the information decay in spectral terms. Pick an arbitrarily small constant $c > 0$ and assume that

$$\|A_s\|_2 \le \frac{\delta}{\|\bar{\mu} - \mu_0\|_2} \Big(\frac{c}{1+c}\Big) \Big(\frac{1}{s}\Big)^{1+c}. \tag{24}$$

By (22), $\mathbb{E}\,\mu_t - \mu_0 = (I_d - Q_t)(\bar{\mu} - \mu_0)$; therefore, by (23),

$$\begin{aligned}
\|\mathbb{E}\,\mu_t - \mu_0\|_2 &\le \|\bar{\mu} - \mu_0\|_2 \sum_{s=1}^{t} \|A_s\|_2 \le \frac{\delta c}{1+c} \sum_{s=1}^{t} s^{-1-c} \\
&\le \frac{\delta c}{1+c} \Big(1 + \int_1^{\infty} x^{-1-c}\,dx\Big) = \delta,
\end{aligned} \tag{25}$$

The relation says that, on average, the means of any of the agents' posteriors can be brought as close to the original mean to be learned as we want. We can turn this into a high-probability event by using some basic random matrix theory. Recall that $\mathbb{E}\,\mu_t$ is itself a random variable whose stochasticity comes from the matrices $X_s$, which are all drawn from Gaussians. Because $M_s$ is positive semidefinite,

$$\|A_s\|_2 \le \|M_s^{-1}\|_2 \le \frac{(\sigma/\bar\sigma)^2}{\lambda_{\min}(X_t^T X_t)} \le \Big(\frac{\sigma/\bar\sigma}{\sigma_1(X_t)}\Big)^2, \tag{26}$$

which gives us a relation between the spectral norm of $(I_s + M_s)^{-1}$ and the smallest singular value $\sigma_1(X_t)$ of an $m_t$-by-$d$ matrix $X_t$ whose elements are drawn *iid* from $N(0,1)$. The asymptotic behavior of $\sigma_1(X_t)$ for large values of $m_t$ has been extensively studied within the field of random matrix theory (Davidson and Szarek, 2001; Edelman, 1988; Rudelson and Vershynin, 2009). Following Theorem II.13 in (Davidson & Szarek (Davidson and Szarek, 2001)), for any $\gamma_t > 0$,

$$\mathbb{P}[\sigma_1(X_t) < \sqrt{m_t} - \sqrt{d} - \gamma_t] \le e^{-\gamma_t^2/2}.$$

We use $C$ below as a generic constant large enough to satisfy the inequalities where it appears. Setting $\gamma_t = C\sqrt{\log((t+1)/\varepsilon)}$ ensures that $\sum_{t>0} e^{-\gamma_t^2/2} < \varepsilon$, hence that $\sigma_1(X_t) < \sqrt{m_t} - \sqrt{d} - \gamma_t$ holds for all $t$ with probability less than $\varepsilon$. With our setting of $m_t$, this means that, for all $t > 0$,

$$\mathbb{P}\Big[\,\sigma_1(X_t) \ge \frac{\sqrt{m_t}}{2}\,\Big] > 1 - \varepsilon. \tag{27}$$

Assuming the event in (27), it follows from (26) and our setting of $m_t$ that

$$\|A_t\|_2 \le \frac{4}{m_t}\Big(\frac{\sigma}{\bar\sigma}\Big)^2 \le \frac{\delta}{\|\bar\mu - \mu_0\|_2}\Big(\frac{4}{D_c}\Big)\Big(\frac{1}{t}\Big)^{1+c};$$

hence (24) holds for $D_c$ large enough. By (25, 27), this proves that, with probability greater than $1-\varepsilon$, $\|\mathbb{E}\,\mu_t - \mu_0\|_2 \le \delta$ for all $t > 0$, which completes the proof. ■

## 4. Hopped Learning

In this section, we consider the "hopped learning" scenario in which learner $t$ hops back to pick a teacher from $\{0, 1, \ldots, t-1\}$ at random, and then takes $m_t$ samples from her posterior. To model the data generating process, we continue to adopt the Gaussian setting from Sections 3.5 and 3.6. Note that the graph sequence $G_t$ becomes random under the constraint that a learner can only get data from the learners before her. Since multiple samples can be sent from a teacher to a learner, we use $d_{t,s,i}$ to denote the $i$-th sample generated by agent $s$ at time $t$. Note that though a learner only receives data from one teacher, without loss of generality, we assume all her predecessors generate samples but she only listens to one of them. The recursive relation for $\mu_t$ becomes

$$\mu_t = \frac{\beta_t}{m_t}\sum_{s=0}^{t-1}\chi_{t,s}\sum_{i=1}^{m_t} d_{t,s,i} + (1-\beta_t)\bar\mu, \tag{28}$$

where, given $t$, the random variable $\chi_{t,s}$ is 1 for a value of $s$ picked at random between 0 and $s-1$, and is zero elsewhere; recall that $\beta_t := m_t\tau/(\bar\tau + m_t\tau)$. Hopped iterated learning provides access

to earlier data, so one would expect the lengths of the training sessions to grow more slowly than in chained learning. The change is indeed quite dramatic:

**Theorem 5** *For any positive $\varepsilon < |\mu_0 - \bar{\mu}|$, the following sample size sequence makes hopped iterating learning $\varepsilon$-self-sustaining:*

$$m_t = B_c \frac{|\mu_0 - \bar{\mu}|}{\varepsilon} \Big(\frac{\sigma}{\bar{\sigma}}\Big)^2 (1 + \log t)^{1+c},$$

*for an arbitrarily small $c > 0$ and a constant $B_c$ that depends only on $c$.*

**Proof** By taking expectation on both sides of (28), for any $t > 0$,

$$\mathbb{E}\,\mu_t = \frac{\beta_t}{t} \sum_{s=0}^{t-1} \mathbb{E}\,\mu_s + (1 - \beta_t)\bar{\mu},$$

We define $\gamma_1 = \beta_1$ and, for $t > 1$,

$$\gamma_t := (1 + \beta_1) \left(1 + \frac{\beta_2}{2}\right) \cdots \left(1 + \frac{\beta_{t-1}}{t-1}\right) \frac{\beta_t}{t}.$$

We verify easily that $\mathbb{E}\,\mu_t = \gamma_t \mu_0 + (1 - \gamma_t)\bar{\mu}$, for $t > 0$; therefore, the first part in establishing $\varepsilon$-self-sustainability consists of proving that

$$1 \geq \gamma_t \geq 1 - \frac{\varepsilon}{|\mu_0 - \bar{\mu}|}\,, \tag{29}$$

which will show that $|\mathbb{E}\,\mu_t - \mu_0| \leq \varepsilon$. Note that

$$\gamma_t \leq \frac{1}{t} \prod_{s=1}^{t-1} \Big(1 + \frac{1}{s}\Big) = 1.$$

Now define

$$\alpha_s = \frac{\varepsilon}{B_c |\mu_0 - \bar{\mu}| s (1 + \log s)^{1+c}}.$$

for $s > 0$. We pick a constant $B_c$ large enough so that $\alpha_s$ is small enough to carry out first-order Taylor approximations around $1 + \alpha_s$. We find that

$$\begin{aligned} 1 + \frac{\beta_s}{s} = 1 + \frac{1}{s}\Big(1 - \frac{1}{1 + m_s \tau/\bar{\tau}}\Big) &\geq \Big(1 + \frac{1}{s}\Big)\Big(1 - \frac{1}{(s+1) m_s \tau/\bar{\tau}}\Big) \\ &\geq \Big(1 + \frac{1}{s}\Big)\Big(1 - \frac{s\alpha_s}{s+1}\Big) \geq \Big(1 + \frac{1}{s}\Big)(1 - \alpha_s) \geq \Big(1 + \frac{1}{s}\Big) e^{-2\alpha_s}. \end{aligned}$$

Thus,

$$\gamma_t \geq \frac{\beta_t}{t} \prod_{s=1}^{t-1} \Big(1 + \frac{1}{s}\Big) e^{-2\sum_{s=1}^{t-1} \alpha_s} = \beta_t e^{-2\sum_{s=1}^{t-1} \alpha_s} \geq 1 - \frac{\varepsilon}{|\mu_0 - \bar{\mu}|},$$

which establishes (29). Our derivation relies on the fact that

$$\beta_t \geq 1 - \frac{\varepsilon}{B_c |\mu_0 - \bar{\mu}| (1 + \log t)^{1+c}} \geq 1 - \frac{\varepsilon}{2|\mu_0 - \bar{\mu}|}$$

and

$$\sum_{s=1}^{t-1} \frac{1}{s(1+\log s)^{1+c}} \le 1 + \frac{1}{(\log e)^{1+c}} \int_2^{t-1} \frac{1}{x(\ln x)^{1+c}}\, dx = O\Big(\frac{1}{c}\Big);$$

hence,

$$e^{-2\sum_{s=1}^{t-1}\alpha_s} \ge e^{-O(\varepsilon/(cB_c|\mu_0-\bar\mu|))} \ge 1 - \frac{\varepsilon}{2|\mu_0-\bar\mu|}.$$

Having shown that $|\mathbb{E}\,\mu_t - \mu_0| \le \varepsilon$ for all $t$, it now suffices to prove that $\sigma_t^2 + \operatorname{var}\mu_t$ remains bounded. We note that $\tau_t > m_t\tau \to \infty$, hence $\sigma_t^2 = 1/\tau_t \to 0$, so the remainder of the proof needs to establish that the variance of $\mu_t$ stays bounded. Writing $D_{t,s} := d_{t,s,1} + \cdots + d_{t,s,m_t}$, we have $\operatorname{var} D_{t,s} = m_t \operatorname{var} d_{t,s,1} = m_t(\sigma_s^2 + \sigma^2 + \operatorname{var}\mu_s)$; hence

$$\mathbb{E}\, D_{t,s}^2 = \operatorname{var} D_{t,s} + (\mathbb{E}\, D_{t,s})^2 = m_t(\sigma_s^2 + \sigma^2 + \operatorname{var}\mu_s) + m_t^2(\mathbb{E}\,\mu_s)^2.$$

In (28), the variables $\chi_{t,s}$ and $D_{t,s}$ are independent, for $0 \le s \le t-1$; furthermore, $\mathbb{E}\,\chi_{t,s} = \mathbb{E}\,\chi_{t,s}^2 = 1/t$, and $\mathbb{E}\,\chi_{t,s_1}\chi_{t,s_2} = 0$ if $s_1 \ne s_2$; therefore,

$$\begin{aligned}\operatorname{var}[\chi_{t,s}D_{t,s}] &= \mathbb{E}\,\chi_{t,s}^2\,\mathbb{E}\,D_{t,s}^2 - (\mathbb{E}\,\chi_{t,s})^2(\mathbb{E}\,D_{t,s})^2 = \frac{\mathbb{E}\,D_{t,s}^2}{t} - \frac{(\mathbb{E}\,D_{t,s})^2}{t^2} \\ &= \Big(\frac{m_t}{t}\Big)\big(\sigma_s^2 + \sigma^2 + \operatorname{var}\mu_s + m_t(\mathbb{E}\,\mu_s)^2\big) - \Big(\frac{m_t}{t}\Big)^2(\mathbb{E}\,\mu_s)^2\end{aligned} \tag{30}$$

and, for $s_1 \ne s_2$,

$$\begin{aligned}\operatorname{cov}[\chi_{t,s_1}D_{t,s_1}, \chi_{t,s_2}D_{t,s_2}] &= \mathbb{E}\,[\chi_{t,s_1}\chi_{t,s_2}D_{t,s_1}, D_{t,s_2}] - \mathbb{E}\,[\chi_{t,s_1}D_{t,s_1}]\mathbb{E}\,[\chi_{t,s_2}D_{t,s_2}] \\ &= \mathbb{E}\,[\chi_{t,s_1}\chi_{t,s_2}]\mathbb{E}\,[D_{t,s_1}D_{t,s_2}] - \mathbb{E}\,\chi_{t,s_1}\mathbb{E}\,D_{t,s_1}\mathbb{E}\,\chi_{t,s_2}\mathbb{E}\,D_{t,s_2} \\ &= -\frac{1}{t^2}\,\mathbb{E}\,D_{t,s_1}\mathbb{E}\,D_{t,s_2} = -\Big(\frac{m_t}{t}\Big)^2\mathbb{E}\,\mu_{s_1}\mathbb{E}\,\mu_{s_2}.\end{aligned} \tag{31}$$

Then, by taking the variance on both sides of (28), we have

$$\begin{aligned}\operatorname{var}\mu_t &= \Big(\frac{\beta_t}{m_t}\Big)^2 \operatorname{var}\sum_{s=0}^{t-1}\chi_{t,s}D_{t,s} \\ &= \Big(\frac{\beta_t}{m_t}\Big)^2\Big(\sum_{s=0}^{t-1}\operatorname{var}[\chi_{t,s}D_{t,s}] + \sum_{0\le s_1\ne s_2\le t-1}\operatorname{cov}[\chi_{t,s_1}D_{t,s_1}, \chi_{t,s_2}D_{t,s_2}]\Big) \\ &= \Big(\frac{\beta_t}{m_t}\Big)^2\Big(\sum_{s=0}^{t-1}\Big(\frac{m_t}{t}\Big)\big(\sigma_s^2 + \sigma^2 + \operatorname{var}\mu_s + m_t(\mathbb{E}\,\mu_s)^2\big) - \Big(\frac{m_t}{t}\Big)^2\Big(\sum_{s=0}^{t-1}\mathbb{E}\,\mu_s\Big)^2\Big) \\ &\le \frac{1}{tm_t}\sum_{s=0}^{t-1}\big(\sigma_s^2 + \sigma^2 + \operatorname{var}\mu_s + m_t(\mathbb{E}\,\mu_s)^2\big).\end{aligned}$$

Notice that $\sigma_s^2 \to 0$ and $(\mathbb{E}\,\mu_s)^2$ is bounded since $|\mathbb{E}\,\mu_t - \mu_0| \le \varepsilon$. We conclude that $\sigma_t^2 + \operatorname{var}\mu_t$ remains bounded for all $t$. ■

## 5. Networked Learning

In this section, we study the information transfer and iterated learning with general graph sequence $G_t$. We assume that the initial belief $\mu_{0,i}$ of agent $i$ is Gaussian: $\mu_{0,i} \sim \mathcal{N}(x_{0,i}, \sigma_{0,i}^2)$. Without loss of generality, the truth is assumed to be a constant (single-point distribution: $\mu_{t,0} = 0$; $\sigma_{t,0} = 0$ for all $t$) and the standard deviation is the same for all other agents, i.e., $\sigma_{0,i} = \sigma_0 > 0$ for $i > 0$. Because agent 0 holds the truth, no edge points out of it. The adjacency matrix of $G_t$ is denoted by $A_t$: it is an $(n+1) \times (n+1)$ matrix whose first row is $(1, 0, \ldots, 0)$. Note that $n$ is the number of learners and should not be confused with the number of hypotheses in Section 3.

### 5.1. The dynamics in matrix form

Let $D_t$ and $P_t$ denote the $(n+1)$-by-$(n+1)$ diagonal matrices $\operatorname{diag}(\eta_{t,i})$ and $(\tau_0/\tau)I + \sum_{k=0}^{t-1} D_k$, respectively, where $\eta_{t,i}$ is the out-degree of agent $i$ at time $t$, $I$ is the identity matrix and the sum is 0 for $t = 0$. It follows from (1) that $\mu_{t,i} \sim \mathcal{N}(x_{t,i}, (\tau P_t)_{ii}^{-1})$ for $i > 0$. Regrouping the means in vector form, $\boldsymbol{x}_t := (x_{t,0}, \ldots, x_{t,n})^T$, where $x_{t,0} = 0$ and $x_{0,1}, \ldots, x_{0,n}$ are given as inputs, we have

$$\boldsymbol{x}_{t+1} = (P_t + D_t)^{-1} \left(P_t \boldsymbol{x}_t + A_t \left(\boldsymbol{x}_t + \boldsymbol{u}_t + \boldsymbol{\varepsilon}_t\right)\right), \tag{32}$$

where $\boldsymbol{u}_t$ is such that $u_{t,0} \sim \mathcal{N}(0, 0)$ and, for $i > 0$, $u_{t,i} \sim \mathcal{N}(\mathbf{0}, (\tau (P_t)_{ii})^{-1})$; and $\boldsymbol{\varepsilon}_t$ is such that $\varepsilon_{t,0} \sim \mathcal{N}(0, 0)$ and, for $i > 0$, $\varepsilon_{t,i} \sim \mathcal{N}(\mathbf{0}, 1/\tau)$. We refer to the vectors $\boldsymbol{x}_t$ and $\boldsymbol{y}_t := \mathbb{E}\, \boldsymbol{x}_t$ as the *mean process* and the *expected mean process*, respectively. Taking expectations on both sides of (32) with respect to the random vectors $\boldsymbol{u}_t$ and $\boldsymbol{\varepsilon}_t$ yields the update rule for the expected mean process: $\boldsymbol{y}_0 = \boldsymbol{x}_0$ and, for $t > 0$,

$$\boldsymbol{y}_{t+1} = (P_t + D_t)^{-1} (P_t + A_t)\, \boldsymbol{y}_t. \tag{33}$$

A key observation is that $(P_t + D_t)^{-1} (P_t + A_t)$ is a stochastic matrix, so the expected mean process $\boldsymbol{y}_t$ forms a diffusive influence system (Chazelle, 2015): the vector evolves by taking convex combinations of its own coordinates. What makes the analysis different from standard multiagent agreement systems is that the weights vary over time. In fact, some weights typically tend to 0, which violates one of the cardinal assumptions used in the analysis of averaging systems (Chazelle, 2015; Moreau, 2005). This leads us to the use of arguments, such as fourth-order moment bounds, that are not commonly encountered in this area.

### 5.2. The results

The belief vector $\boldsymbol{\mu}_t$ is Gaussian with mean $\boldsymbol{x}_t$ and covariance matrix $\Sigma_t$ formed by zeroing out the top-left element of $(\tau P_t)^{-1}$. We say that the system reaches *truthful consensus* if both the mean process $\boldsymbol{x}_t$ and the covariance matrix tend to zero as $t$ goes to infinity. This indicates that all the agents' beliefs share a common mean equal to the truth and the "error bars" vanish over time. In view of (1), the covariance matrix indeed tends to 0 as long as the degrees are nonzero infinitely often, a trivial condition. To establish truthful consensus, therefore, boils down to studying the mean process $\boldsymbol{x}_t$. We do this in two parts: first, we show that the expected mean process converges to the truth; then we prove that fluctuations around it eventually vanish almost surely.[1]

1. The Kullback-Leibler divergence (Jadbabaie et al., 2012) is not suitable here because the estimator is Gaussian, hence continuous, whereas the truth is a single-point distribution.

*Truth-hearing assumption*: Given any interval of length $\kappa := \lfloor 1/\gamma \rfloor$, every agent $i > 0$ has an edge $(i, 0)$ in $G_t$ for at least one value of $t$ in that interval.

**Theorem 6** *Under the truth-hearing assumption, the system reaches truthful consensus with a convergence rate bounded by $O(t^{-\gamma/2\eta})$, where $\eta$ is the maximum outdegree over all the networks.*

We prove the theorem in the next two sections. It will follow directly from Lemmas 7 and 8 below. The convergence rate can be improved to the order of $t^{-(1-\varepsilon)\gamma/\eta}$, for arbitrarily small $\varepsilon > 0$. The inverse dependency on $\gamma$ is not surprising: the more access to the truth the stronger the attraction to it. On the other hand, it might seem counterintuitive that a larger outdegree should slow down convergence. This illustrates the risk of groupthink. It pays to follow the crowds when the crowds are right. When they are not, however, this distracts from the lonely voice that happens to be right.

How essential is the truth-hearing assumption? We show that it is necessary. Simply having access to the truth infinitely often is not enough to achieve truthful consensus.

### 5.3. The proofs

In this subsection, we demonstrate technical details for the proof of the results. We begin with some repeated used inequalities.

#### 5.3.1. USEFUL MATRIX INEQUALITIES

We highlight certain matrix inequalities to be used throughout. We use the standard element-wise notation $R \leq S$ to indicate that $R_{ij} \leq S_{ij}$ for all $i, j$. The infinity norm $\|R\|_\infty = \max_i \sum_j |r_{ij}|$ is submultiplicative: $\|RS\|_\infty \leq \|R\|_\infty \|S\|_\infty$, for any matching rectangular matrices. On the other hand, the max-norm $\|R\|_{\max} := \max_{i,j} |r_{ij}|$ is not, but it is transpose-invariant and also satisfies: $\|RS\|_{\max} \leq \|R\|_\infty \|S\|_{\max}$. It follows that

$$\begin{aligned} \|RSR^T\|_{\max} &\leq \|R\|_\infty \|SR^T\|_{\max} = \|R\|_\infty \|RS^T\|_{\max} \\ &\leq \|R\|_\infty^2 \|S^T\|_{\max} = \|R\|_\infty^2 \|S\|_{\max}. \end{aligned} \tag{34}$$

#### 5.3.2. THE EXPECTED MEAN PROCESS DYNAMICS

We analyze the convergence of the mean process in expectation. The expected mean $\boldsymbol{y}_t = \mathbb{E}\, \boldsymbol{x}_t$ evolves through an averaging process entirely determined by the initial value $\boldsymbol{y}_0 = (0, x_{0,1}, \ldots, x_{0,n})^T$ and the graph sequence $G_t$. Intuitively, if an agent communicates repeatedly with a holder of the truth, the weight of the latter should accumulate and increasingy influence the belief of the agent in question. Our goal in this section is to prove the following result:

**Lemma 7** *Under the truth-hearing assumption, the expected mean process $\boldsymbol{y}_t$ converges to the truth asymptotically. If, at each step, no agent receives information from more than $\eta$ agents, then the convergence rate is bounded by $Ct^{-\gamma/2\eta}$, where $C$ is a constant that depends on $\boldsymbol{x}_0, \gamma, \eta, \sigma_0/\sigma$.*

**Proof** We define $B_t$ as the matrix formed by removing the first row and the first column from the stochastic $P_{t+1}^{-1}(P_t + A_t)$. If we write $\boldsymbol{y}_t$ as $(0, \boldsymbol{z}_t)$ then, by (33),

$$\begin{pmatrix} 0 \\ \boldsymbol{z}_{t+1} \end{pmatrix} = \begin{pmatrix} 1 & \mathbf{0} \\ \boldsymbol{\alpha}_t & B_t \end{pmatrix} \begin{pmatrix} 0 \\ \boldsymbol{z}_t \end{pmatrix}, \tag{35}$$

where $\boldsymbol{\alpha}_{t,i} = (P_{t+1}^{-1})_{ii}$ if there is an edge $(i, 0)$ at time $t$ and $\boldsymbol{\alpha}_{t,i} = 0$ otherwise. This further simplifies to

$$\boldsymbol{z}_{t+1} = B_t \boldsymbol{z}_t. \tag{36}$$

Let $\mathbf{1}$ be the all-one column vector of length $n$. Since $P_{t+1}^{-1}(P_t + A_t)$ is stochastic,

$$\boldsymbol{\alpha}_t + B_t \mathbf{1} = \mathbf{1} \tag{37}$$

In matrix terms, the truth-hearing assumption means that, for any $t \geq 0$,

$$\boldsymbol{\alpha}_t + \boldsymbol{\alpha}_{t+1} + \cdots + \boldsymbol{\alpha}_{t+\kappa-1} \geq Q_{t+\kappa}^{-1} \mathbf{1}, \tag{38}$$

where $Q_t$ is the matrix derived from $P_t$ by removing the first row and the last column; the inequality relies on the fact that $P_t$ is monotonically nondecreasing. For any $t > s \geq 0$, we define the product matrix $B_{t:s}$ defined as

$$B_{t:s} := B_{t-1} B_{t-2} \ldots B_s, \tag{39}$$

with $B_{t:t} = I$. By (36), for any $t > s \geq 0$,

$$\boldsymbol{z}_t = B_{t:s}\, \boldsymbol{z}_s. \tag{40}$$

To bound the infinity norm of $B_{t:0}$, we observe that, for any $0 \leq l < \kappa - 1$, the $i$-th diagonal element of $B_{s+\kappa:s+l+1}$ is lower-bounded by

$$\begin{aligned} \prod_{j=l+1}^{\kappa-1} (B_{s+j})_{ii} &= \prod_{j=l+1}^{\kappa-1} \frac{(P_{s+j} + A_{s+j})_{ii}}{(P_{s+j+1})_{ii}} \\ \geq \prod_{j=l+1}^{\kappa-1} \frac{(P_{s+j})_{ii}}{(P_{s+j+1})_{ii}} &= \frac{(P_{s+l+1})_{ii}}{(P_{s+\kappa})_{ii}} \geq \frac{(P_s)_{ii}}{(P_{s+\kappa})_{ii}}. \end{aligned} \tag{41}$$

The inequalities follow from the nonnegativity of the entries and the monotonicity of $(P_t)_{ii}$. Note that (41) also holds for $l = \kappa - 1$ since $(B_{s+\kappa:s+\kappa})_{ii} = 1$.

Since $P_{t+1}^{-1}(P_t + A_t)$ is stochastic, the row-sum of $B_t$ does not exceed 1; therefore, by pre-multiplying $B_{s+1}, B_{s+2}, \ldots$ on both sides of (37), we obtain:

$$B_{s+\kappa:s} \mathbf{1} \leq \mathbf{1} - \sum_{l=0}^{\kappa-1} B_{s+\kappa:s+l+1} \boldsymbol{\alpha}_{s+l}. \tag{42}$$

Noting that $\|B_t\|_\infty = \|B_t \mathbf{1}\|_\infty$ for any $t$, as $B_t$ is non-negative, we combine (38), (41), and (42) together to derive:

$$\|B_{s+\kappa:s}\|_\infty \leq 1 - \min_{i>0} \frac{(P_s)_{ii}}{(P_{s+\kappa})_{ii}^2}. \tag{43}$$

Let $\eta := \max_{t\geq 0} \max_{1\leq i\leq n} \eta_{t,i}$ denote the maximum outdegree in all the networks, and define $\delta = \min\{\tau_0/\tau, 1\}$. For any $i > 0$ and $s \geq \kappa$,

$$\frac{s\delta}{\kappa} \leq (P_s)_{ii} \leq \eta s + \frac{\tau_0}{\tau}; \tag{44}$$

hence,

$$\max_i (P_{s+\kappa})_{ii} \leq \eta(s+\kappa) + \frac{\tau_0}{\tau}. \tag{45}$$

It follows that

$$\frac{(P_{s+\kappa})_{ii} - (P_s)_{ii}}{(P_{s+\kappa})_{ii}} = \frac{\sum_{l=0}^{\kappa-1} \eta_{s+l,i}}{(P_{s+\kappa})_{ii}} \leq \frac{\eta\kappa^2\delta^{-1}}{s+\kappa}. \tag{46}$$

Thus, we have

$$\begin{aligned} \min_{i>0} \frac{(P_s)_{ii}}{(P_{s+\kappa})_{ii}} &= 1 - \max_{i>0} \frac{(P_{s+\kappa})_{ii} - (P_s)_{ii}}{(P_{s+\kappa})ii} \\ &\geq 1 - \frac{\eta\kappa^2\delta^{-1}}{s+\kappa}. \end{aligned} \tag{47}$$

We can replace the upper bound of (43) by

$$1 - \frac{1}{\max_{i>0}(P_{s+\kappa})_{ii}} \min_{i>0} \frac{(P_s)_{ii}}{(P_{s+\kappa})_{ii}^2},$$

which, together with (45) and (47) gives us

$$\begin{aligned} \|B_{s+\kappa:s}\|_\infty &\leq 1 - \frac{1}{\eta(s+\kappa) + \tau_0/\tau} \left(1 - \frac{\eta\kappa^2\delta^{-1}}{s+\kappa}\right) \\ &\leq 1 - \frac{1}{2\eta\kappa(m+2)}. \end{aligned} \tag{48}$$

The latter inequality holds as long as $s = m\kappa > 0$ and

$$m \geq m^* := \frac{2\eta\kappa}{\delta} + \frac{\tau_0}{\eta\kappa\tau}.$$

It follows that, for $m_0 \geq m^*$,

$$\begin{aligned} \|B_{(m_0+m)\kappa:m_0\kappa}\|_\infty &\leq \prod_{j=2}^{m+1} \left(1 - \frac{1}{2\eta\kappa(m_0+j)}\right) \\ &\leq \exp\left\{-\frac{1}{2\eta\kappa} \sum_{j=2}^{m+1} \frac{1}{m_0+j}\right\}. \end{aligned} \tag{49}$$

The matrices $B_t$ are sub-stochastic so that

$$\|B_t\, \boldsymbol{z}\|_\infty \leq \|B_t\|_\infty \|\boldsymbol{z}\|_\infty \leq \|\boldsymbol{z}\|_\infty.$$

By (40), for any $t \geq (m_0+m)\kappa$,

$$\boldsymbol{z}_t = B_{t:(m_0+m)\kappa} B_{(m_0+m)\kappa:m_0\kappa}\, \boldsymbol{z}_{m_0},$$

so that, by using standard bounds for the harmonic series, $\ln(k+1) < 1 + \frac{1}{2} + \cdots + \frac{1}{k} \leq 1 + \ln k$, we find that

$$\begin{aligned}\|\boldsymbol{z}_t\|_\infty &\leq \|B_{(m_0+m)\kappa:m_0\kappa}\, \boldsymbol{z}_{m_0}\|_\infty \\ &\leq \|B_{(m_0+m)\kappa:m_0\kappa}\|_\infty \|\boldsymbol{z}_0\|_\infty \\ &\leq C t^{-1/(2\eta\kappa)},\end{aligned}$$

where $C > 0$ depends on $\boldsymbol{z}_0, \kappa, \eta, \tau_0/\tau$. We note that the convergence rate can be improved to the order of $t^{-(1-\varepsilon)\gamma/\eta}$, for arbitrarily small $\varepsilon > 0$, by working a little harder with (48). ■

### 5.3.3. THE MEAN PROCESS DYNAMICS

Recall that $\mu_{t,i} \sim \mathcal{N}(x_{t,i}, \tau_{t,i}^{-1})$, where $\tau_{t,i}$ denotes the precision $\sigma_{t,i}^{-2}$. A key observation about the updating rule in (1) is that the precision $\tau_{t,i}$ is entirely determined by the graph sequence $G_t$ and is independent of the actual dynamics. Adding to this the connectivity property implied by the truth-hearing assumption, we find immediately that $\tau_{t,i} \to \infty$ for any agent $i$. This ensures that the covariance matrix $\Sigma_t$ tends to 0 as $t$ goes to infinity, which satisfies the second criterion for truthful consensus. The first criterion requires that the mean process $\boldsymbol{x}_t$ should converge to the truth $\mathbf{0}$. Take the vector $\boldsymbol{x}_t - \boldsymbol{y}_t$ and remove the first coordinate $(\boldsymbol{x}_t - \boldsymbol{y}_t)_0$ to form the vector $\boldsymbol{\Delta}_t \in \mathbb{R}^n$. Under the truth-hearing assumption, we have seen that $\boldsymbol{y}_t \to \mathbf{0}$ (Lemma 7), so it suffices to prove the following:

**Lemma 8** *Under the truth-hearing assumption, the deviation $\boldsymbol{\Delta}_t$ vanishes almost surely.*

**Proof** We use a fourth-moment argument. The justification for the high order is technical: it is necessary to make a certain "deviation power" series converge. By (32), $\boldsymbol{x}_t$ is a linear combination of independent Gaussian random vectors $\boldsymbol{u}_s$ and $\boldsymbol{\varepsilon}_s$ for $0 \leq s \leq t-1$, and thus $\boldsymbol{x}_t$ itself is a Gaussian random vector. Therefore $\boldsymbol{\Delta}_t$ is also Gaussian and its mean is zero. From Markov's inequality, for any $c > 0$,

$$\sum_{t\geq 0} \mathbb{P}[|\Delta_{t,i}| \geq c] \leq \sum_{t\geq 0} \frac{\mathbb{E}\,\Delta_{t,i}^4}{c^4}. \tag{50}$$

If we are able to show the right hand side of (50) is finite for any $c > 0$, then, by the Borel-Cantelli lemma, with probability one, the event $|\Delta_{t,i}| \geq c$ occurs only a finite number of times, and so $\Delta_{t,i}$ goes to zero almost surely. Therefore, we only need to analyze the order of the fourth moment $\mathbb{E}\,\Delta_{t,i}^4$. By subtracting (33) from (32), we have:

$$\boldsymbol{\Delta}_{t+1} = B_t \boldsymbol{\Delta}_t + M_t \boldsymbol{v}_t, \tag{51}$$

where $\boldsymbol{v}_t := \boldsymbol{u}_t + \boldsymbol{\varepsilon}_t$ and $M_t := P_{t+1}^{-1} A_t$; actually, for dimensions to match, we remove the top coordinate of $\boldsymbol{v}_t$ and the first row and first column of $M_t$ (see previous section for definition of $B_t$). Transforming the previous identity into a telescoping sum, it follows from $\boldsymbol{\Delta}_0 = \boldsymbol{x}_0 - \boldsymbol{y}_0 = \mathbf{0}$ and the definition $B_{t:s} = B_{t-1} B_{t-2} \ldots B_s$ that

$$\boldsymbol{\Delta}_t = \sum_{s=0}^{t-1} B_{t:s+1} M_s \boldsymbol{v}_s = \sum_{s=0}^{t-1} R_{t,s} \boldsymbol{v}_s, \tag{52}$$

where $R_{t,s} := B_{t:s+1}M_s$. We denote by $C_1, C_2, \ldots$ suitably large constants (possibly depending on $\kappa, \eta, n, \tau, \tau_0$). By (44), $\|M_s\|_\infty \le C_1/(s+1)$ and, by (49), for sufficiently large $s$,

$$\|B_{t:s+1}\|_\infty \le C_2(s+1)^\beta (t+1)^{-\beta},$$

where $\beta = 1/2\eta\kappa < 1$. Combining the above inequalities, we obtain the following estimate of $R_{t,s}$ as

$$\|R_{t,s}\|_\infty \le C_3(s+1)^{-1+\beta}(t+1)^{-\beta}. \tag{53}$$

In the remainder of the proof, the power of a vector is understood element-wise. We use the fact that $\boldsymbol{v}_s$ and $\boldsymbol{v}_{s'}$ are independent if $s \neq s'$ and that the expectation of an odd power of an unbiased Gaussian is always zero. By Cauchy-Schwarz and Jensen's inequalities,

$$\begin{aligned}
\mathbb{E}\,\boldsymbol{\Delta}_t^4 &= \left(\sum_{s=0}^{t-1} R_{t,s}\boldsymbol{v}_s\right)^4 \\
&= \sum_{s=0}^{t-1} \mathbb{E}(R_{t,s}\boldsymbol{v}_s)^4 + \sum_{0\le s\neq s'<t} 3\,\mathbb{E}(R_{t,s}\boldsymbol{v}_s)^2\mathbb{E}(R_{t,s'}\boldsymbol{v}_{s'})^2 \\
&\le \sum_{s=0}^{t-1} \mathbb{E}(R_{t,s}\boldsymbol{v}_s)^4 + 3\left(\sum_{s=0}^{t-1} \mathbb{E}(R_{t,s}\boldsymbol{v}_s)^2\right)^2 \\
&\le \sum_{s=0}^{t-1} \mathbb{E}(R_{t,s}\boldsymbol{v}_s)^4 + 3t\sum_{s=0}^{t-1} \mathbb{E}^2(R_{t,s}\boldsymbol{v}_s)^2 \\
&\le (3t+1)\sum_{s=0}^{t-1} \mathbb{E}(R_{t,s}\boldsymbol{v}_s)^4.
\end{aligned} \tag{54}$$

Notice that since the variance of $\boldsymbol{v}_t = (v_{t,1}, \ldots, v_{t,n})^T$ is nonincreasing, there exists a constant $C_4$ such that $\mathbb{E}\, v_{t,i}^4 \le C_4$. By Jensen's inequality and the fact that the variables $v_{t,i}$ are independent for different values of $i$, we have, for any $i, j, k, l$,

$$|\mathbb{E}\, v_{t,i}v_{t,j}v_{t,k}v_{t,l}| \le \max_k \mathbb{E}\, v_{t,k}^4.$$

By direct calculation, it then follows that

$$\begin{aligned}
\max_i \mathbb{E}(R_{t,s}\boldsymbol{v}_s)_i^4 &= \max_i \mathbb{E}\Big(\sum_{j=1}^n (R_{t,s})_{ij}v_{s,j}\Big)^4 \\
&\le \max_i \Big(\sum_{j=1}^n (R_{t,s})_{i,j}\Big)^4 \max_k \mathbb{E}\, v_{s,k}^4 \\
&= \|R_{t,s}\|_\infty^4 \max_k \mathbb{E}\, v_{s,k}^4 \\
&\le C_5(s+1)^{-4+4\beta}(t+1)^{-4\beta}.
\end{aligned} \tag{55}$$

Summing (55) over $0 \le s \le t-1$, we conclude from (54) that $\mathbb{E}\,\boldsymbol{\Delta}_t^4 \le C_6 t^{-2}$, and thus

$$\sum_{t\ge 0} \mathbb{E}\,\boldsymbol{\Delta}_t^4 \le C_6 \sum_{t\ge 1} t^{-2} \le C_7. \tag{56}$$

By the Borel-Cantelli lemma, it follows that $\boldsymbol{\Delta}_t$ vanishes almost surely. ■

Theorem 6 follows directly from Lemmas 7 and 8.

**Why the truth-hearing assumption is necessary.** We describe a sequence of graphs $G_t$ that allows every agent infinite access to the truth and yet does not lead to truthful consensus. For this, it suffices to ensure that the expected mean process $\boldsymbol{y}_t$ does not converge. Consider a system with two learning agents with priors $\mu_{0,1}$ and $\mu_{0,2}$ from the same distribution $\mathcal{N}(2,1)$. We have $x_{0,1} = x_{0,2} = y_{0,1} = y_{0,2} = 2$ and, as usual, the truth is assumed to be 0; the noise variance is $\sigma^2 = 1$. The graph sequence is defined as follows: set $t_1 = 0$; for $k = 1, 2, \ldots$, agent 1 links to the truth agent at time $t_k$ and to agent 2 at times $t_k + 1, \ldots, s_k - 1$; then at time $s_k$, agent 2 links to the truth agent, and then to agent 1 at times $s_k + 1, \ldots, t_{k+1} - 1$. Other than the links mentioned, we assume no additional link exists. The time points $s_k$ and $t_k$ are defined recursively to ensure that

$$y_{s_k,1} \geq 1 + 2^{-2k+1} \quad \text{and} \quad y_{t_k,2} \geq 1 + 2^{-2k}. \tag{57}$$

In this way, the expected mean processes of the two agents alternate while possibly sliding down toward 1 but never lower. The existence of these time points can be proved by induction. Since $y_{0,2} = 2$, the inequality $y_{t_k,2} \geq 1 + 2^{-2k}$ holds for $k = 1$, so let's assume it holds up to $k > 0$. The key to the proof is that, by (33), as agent 1 repeatedly links to agent 2, she is pulled arbitrarily close to it. Indeed, the transition rule gives us

$$y_{t+1,1} = \frac{(P_t)_{11}}{(P_{t+1})_{11}}\, y_{t,1} + \frac{1}{(P_{t+1})_{11}}\, y_{t,2},$$

where $(P_{t+1})_{11} = (P_t)_{11} + 1$, which implies that $y_{t,1}$ can be brought arbitrarily close to $y_{t,2}$ while the latter does not move: this follows from the fact that any product of the form $\prod_{t>t_a}^{t_b} \frac{t}{t+1}$ tends to 0 as $t_b$ grows.[2] Thus a suitably increasing sequence of $s_k, t_k$ ensures the two conditions (57). The beliefs of the two agents do not converge to the truth even though they link to the truth agent infinitely often.

## Acknowledgments

We wish to thank the anonymous referees for their useful comments and suggestions. Some results of this work were previously published in the conferences, Innovations in Theoretical Computer Science (ITCS 2017), and American Control Conference (ACC 2017) by the same authors. Chu Wang did this work prior to joining Amazon. The research of Bernard Chazelle was sponsored by the Army Research Office and the Defense Advanced Research Projects Agency and was accomplished under Grant Number W911NF-17-1-0078. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office, the Defense Advanced Research Projects Agency, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

2. We note that the construction shares a family resemblance with one used by Moreau (Moreau, 2005) to show the non-consensual dynamics of certain multiagent averaging systems. The difference here is that the weights of the averaging change at each step by increasing the agent's self-confidence.

## References

Daron Acemoglu and Asuman Ozdaglar. Opinion dynamics and learning in social networks. *Dynamic Games and Applications*, 1(1):3–49, 2011.

Daron Acemoglu, Munther A Dahleh, Ilan Lobel, and Asuman Ozdaglar. Bayesian learning in social networks. *The Review of Economic Studies*, 78(4):1201–1236, 2011.

Venkatesh Bala and Sanjeev Goyal. Learning from neighbours. *The review of economic studies*, 65 (3):595–621, 1998.

Abhijit V Banerjee. A simple model of herd behavior. *The quarterly journal of economics*, 107(3): 797–817, 1992.

FC Bartlett. Remembering: a study in experimental and social psychology (1932).

Aaron Beppu and Thomas L Griffiths. Iterated learning and the cultural ratchet. In *Proceedings of the 31st annual conference of the cognitive science society*, pages 2089–2094, 2009.

A Bhattachayya. On a measure of divergence between two statistical population defined by their population distributions. *Bulletin Calcutta Mathematical Society*, 35:99–109, 1943.

George E.P. Box and George C. Tiao. *Bayesian inference in statistical analysis*, volume 40. John Wiley & Sons, 2011.

Bernard Chazelle. Diffusive influence systems. *SIAM Journal on Computing*, 44(5):1403–1442, 2015.

Bernard Chazelle and Chu Wang. Inertial Hegselmann-Krause systems. In *Proceedings of the IEEE American Control Conference (ACC)*, pages 1936–1941, 2016.

Dorin Comaniciu, Visvanathan Ramesh, and Peter Meer. Real-time tracking of non-rigid objects using mean shift. In *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, volume 2, pages 142–149. IEEE, 2000.

Kenneth R Davidson and Stanislaw J Szarek. Local operator theory, random matrices and banach spaces. *Handbook of the geometry of Banach spaces*, 1(317-366):131, 2001.

Morris H. DeGroot. Reaching a consensus. *Journal of the American Statistical Association*, 69(345): 118–121, 1974.

Alan Edelman. Eigenvalues and condition numbers of random matrices. *SIAM Journal on Matrix Analysis and Applications*, 9(4):543–560, 1988.

Andrew Gelman, Hal S Stern, John B Carlin, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 2013.

Benjamin Golub and Matthew O. Jackson. Naive learning in social networks and the wisdom of crowds. *American Economic Journal: Microeconomics*, 2(1):112–149, 2010.

Benjamin Golub and Matthew O. Jackson. How homophily affects the speed of learning and best response dynamics. 2012.

Thomas L. Griffiths and Michael L. Kalish. A bayesian view of language evolution by iterated learning. In *Proceedings of the 27th annual conference of the cognitive science society*, pages 827–832, 2005.

Thomas L. Griffiths and Michael L. Kalish. Language evolution by iterated learning with bayesian agents. *Cognitive Science*, 31(3):441–480, 2007.

Thomas L Griffiths, Michael L Kalish, and Stephan Lewandowsky. Theoretical and empirical evidence for the impact of inductive biases on cultural evolution. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 363(1509):3503–3514, 2008.

Michiel Hazewinkel. *Encyclopaedia of Mathematics*. Springer Science & Business Media, 2013.

Rainer Hegselmann and Ulrich Krause. Opinion dynamics and bounded confidence models, analysis, and simulation. *Journal of Artificial Societies and Social Simulation*, 5(3), 2002.

Ali Jadbabaie, Pooya Molavi, Alvaro Sandroni, and Alireza Tahbaz-Salehi. Non-bayesian social learning. *Games and Economic Behavior*, 76(1):210–225, 2012.

Ali Jadbabaie, Pooya Molavi, and Alireza Tahbaz-Salehi. Information heterogeneity and the speed of learning in social networks. *Columbia Business School Research Paper*, (13-28), 2013.

Michael L Kalish, Thomas L Griffiths, and Stephan Lewandowsky. Iterated learning: Intergenerational knowledge transmission reveals inductive biases. *Psychonomic Bulletin & Review*, 14(2): 288–294, 2007.

Simon Kirby, Tom Griffiths, and Kenny Smith. Iterated learning and the evolution of language. *Current opinion in neurobiology*, 28:108–114, 2014.

Ilan Lobel and Evan Sadler. Preferences, homophily, and social learning. *Operations Research*, 64 (3):564–584, 2015.

Soheil Mohajer and Behrouz Touri. On convergence rate of scalar hegselmann-krause dynamics. In *American Control Conference (ACC), 2013*, pages 206–210. IEEE, 2013.

Pooya Molavi, Alireza Tahbaz-Salehi, and Ali Jadbabaie. Foundations of non-bayesian social learning. *Columbia Business School Research Paper*, 2015.

Luc Moreau. Stability of multiagent systems with time-dependent communication links. *IEEE Transactions on Automatic Control*, 50:169–182, 2005.

Elchanan Mossel, Allan Sly, and Omer Tamuz. From agreement to asymptotic learning. *Arxiv preprint arXiv*, 1105, 2011.

Manuel Mueller-Frank. A general framework for rational learning in social networks. *Theoretical Economics*, 8(1):1–40, 2013.

James R Norris. *Markov chains*. Cambridge university press, 1998.

Amy Perfors and Daniel Navarro. Language evolution is shaped by the structure of the world: An iterated learning analysis. Cognitive Science Society, 2011.

Anna N Rafferty, Thomas L Griffiths, and Dan Klein. Convergence bounds for language evolution by iterated learning. In *Proceedings of the Thirty-First Annual Conference of the Cognitive Science Society*, 2009.

Anna N Rafferty, Thomas L Griffiths, and Dan Klein. Analyzing the rate at which languages lose the influence of a common ancestor. *Cognitive science*, 38(7):1406–1431, 2014.

Mohammad Amin Rahimian and Ali Jadbabaie. Learning without recall from actions of neighbors. In *2016 American Control Conference (ACC)*, pages 1060–1065. IEEE, 2016a.

Mohammad Amin Rahimian and Ali Jadbabaie. Naive social learning in ising networks. In *2016 American Control Conference (ACC)*, pages 1088–1093. IEEE, 2016b.

Mohammad Amin Rahimian, Shahin Shahrampour, and Ali Jadbabaie. Learning without recall by random walks on directed graphs. In *2015 54th IEEE Conference on Decision and Control (CDC)*, pages 5538–5543. IEEE, 2015a.

Mohammad Amin Rahimian et al. Learning without recall: A case for log-linear learning. *IFAC-PapersOnLine*, 48(22):46–51, 2015b.

Mark Rudelson and Roman Vershynin. Smallest singular value of a random rectangular matrix. *Communications on Pure and Applied Mathematics*, 62(12):1707–1739, 2009.

Kenny Smith. Iterated learning in populations of bayesian agents. In *Proceedings of the 31st annual conference of the cognitive science society*, pages 697–702, 2009.

Alireza Tahbaz-Salehi, Alvaro Sandroni, and Ali Jadbabaie. Learning under social influence. In *Decision and Control, 2009 held jointly with the 2009 28th Chinese Control Conference. CDC/CCC 2009. Proceedings of the 48th IEEE Conference on*, pages 1513–1519. IEEE, 2009.

Mónica Tamariz and Simon Kirby. Culture: copying, compression, and conventionality. *Cognitive science*, 39(1):171–183, 2015.

## Appendix A. Analysis of the Root-Sine-Distance

In this appendix, we provide detailed discussion of the root-sine-distance. We will first show that it is a metric, followed by the proof of its equivalency to Euclidean distance and the discussion of its relation to other similar metrics. The notation in the appendix is local and should not be confused with the one used in the main body of this paper.

**The root-sine-distance is a metric.** The two forms of the function $d_{RS}$ in (4) make it clear that $0 \le d_{RS}(\boldsymbol{a},\boldsymbol{b}) \le 1$ and $d_{RS}(\boldsymbol{a},\boldsymbol{b}) = 0$ if and only if $\boldsymbol{a}$ and $\boldsymbol{b}$ are identical. We easily check that $d_{RS}$ makes the simplex $\mathcal{S}$ of distributions over $\mathcal{D}$ into a metric space. Indeed, $d_{RS}(\cdot,\cdot)$ is obviously symmetric, and $d_{RS}(\boldsymbol{a},\boldsymbol{b}) = 0$ implies that $\boldsymbol{a} = \boldsymbol{b}$. To check the triangular inequality, notice that

$$d_{RS}(\boldsymbol{a},\boldsymbol{b}) = \sqrt{1 - \Big(\sum_{i=1}^{s} \sqrt{a_i b_i}\Big)^2} = \sin\langle \sqrt{\boldsymbol{a}}, \sqrt{\boldsymbol{b}}\,\rangle, \tag{58}$$

where $\langle \sqrt{\boldsymbol{a}}, \sqrt{\boldsymbol{b}}\,\rangle$ is the angle between the unit vectors $\sqrt{\boldsymbol{a}}$ and $\sqrt{\boldsymbol{b}}$, using the notation $\sqrt{\boldsymbol{v}} = (\sqrt{v_1}, \ldots, \sqrt{v_s})$. To prove that $d_{RS}(\boldsymbol{a},\boldsymbol{b}) + d_{RS}(\boldsymbol{b},\boldsymbol{c}) \ge d_{RS}(\boldsymbol{a},\boldsymbol{c})$ for any $\boldsymbol{a},\boldsymbol{b},\boldsymbol{c} \in \mathcal{S}$, we denote by $\alpha,\beta,\gamma$ the corresponding angles in that order, ie, $\alpha = \langle \sqrt{\boldsymbol{a}}, \sqrt{\boldsymbol{b}}\,\rangle$, etc. The coordinates in $\boldsymbol{a},\boldsymbol{b},\boldsymbol{c}$ are nonnegative; therefore $0 \le \alpha,\beta,\gamma \le \pi/2$. These form the three angles at the origin of a tetrahedron with a vertex at the origin; therefore, by the triangular inequality in spherical geometry, $\alpha + \beta \ge \gamma$. If $\alpha + \beta \le \frac{\pi}{2}$, then $\sin\alpha + \sin\beta \ge \sin\alpha\cos\beta + \cos\alpha\sin\beta = \sin(\alpha+\beta) \ge \sin\gamma$. On the other hand, if $\alpha+\beta > \pi/2$, then $\sin\alpha + \sin\beta = 2\sin\frac{\alpha+\beta}{2}\cos\frac{\alpha-\beta}{2} \ge 2\sin\frac{\pi}{4}\cos\frac{\pi}{4} = 1 \ge \sin\gamma$, which establishes the triangular inequality.

**Relation to the Euclidean distance.** Shrinking the simplex $\mathcal{S}$ by a tiny amount, we define $\mathcal{S}_\varepsilon := \{\boldsymbol{a} \in \mathcal{S} : \varepsilon \le a_i \le 1-\varepsilon\}$ and note that

$$d_E(\boldsymbol{a},\boldsymbol{b}) := \|\boldsymbol{a}-\boldsymbol{b}\|_2 = \sqrt{\sum_{i=1}^{s} (\sqrt{a_i} - \sqrt{b_i})^2 (\sqrt{a_i} + \sqrt{b_i})^2}.$$

It follows that, for $\boldsymbol{a},\boldsymbol{b} \in \mathcal{S}_\varepsilon$,

$$\frac{1}{2} d_E(\boldsymbol{a},\boldsymbol{b}) \le d_E(\sqrt{\boldsymbol{a}}, \sqrt{\boldsymbol{b}}\,) \le \frac{1}{2\sqrt{\varepsilon}}\, d_E(\boldsymbol{a},\boldsymbol{b}). \tag{59}$$

On the other hand, $\|\sqrt{\boldsymbol{a}}\|_2 = \|\sqrt{\boldsymbol{b}}\|_2 = 1$, so the vectors $\sqrt{\boldsymbol{a}}$ and $\sqrt{\boldsymbol{b}}$ form an isosceles triangle; hence

$$d_E(\sqrt{\boldsymbol{a}}, \sqrt{\boldsymbol{b}}\,) = 2\sin\frac{1}{2}\langle \sqrt{\boldsymbol{a}}, \sqrt{\boldsymbol{b}}\rangle = \frac{\sin\langle \sqrt{\boldsymbol{a}}, \sqrt{\boldsymbol{b}}\rangle}{\cos\frac{1}{2}\langle \sqrt{\boldsymbol{a}}, \sqrt{\boldsymbol{b}}\rangle} = \frac{d_{RS}(\boldsymbol{a},\boldsymbol{b})}{\cos\frac{1}{2}\langle \sqrt{\boldsymbol{a}}, \sqrt{\boldsymbol{b}}\rangle}\,.$$

Since $0 \le \langle \sqrt{\boldsymbol{a}}, \sqrt{\boldsymbol{b}}\,\rangle \le \frac{\pi}{2}$,

$$d_{RS}(\boldsymbol{a},\boldsymbol{b}) \le d_E(\sqrt{\boldsymbol{a}}, \sqrt{\boldsymbol{b}}\,) \le \sqrt{2}\, d_{RS}(\boldsymbol{a},\boldsymbol{b}).$$

Together with (59) this shows that, for any $\boldsymbol{a},\boldsymbol{b} \in \mathcal{S}_\varepsilon$,

$$\frac{1}{2\sqrt{2}}\, d_E(\boldsymbol{a},\boldsymbol{b}) \le d_{RS}(\boldsymbol{a},\boldsymbol{b}) \le \frac{1}{2\sqrt{\varepsilon}}\, d_E(\boldsymbol{a},\boldsymbol{b}), \tag{60}$$

which shows that the Euclidean distance and the metric $d_{RS}$ are equivalent in $\mathcal{S}_\varepsilon$.

**Relation to other distances.** The metric $d_{RS}$ is related to the Hellinger and Bhattacharyya distances. Writing $C(\boldsymbol{a}, \boldsymbol{b}) = \sum_{i=1}^{s} \sqrt{a_i b_i}$ (Comaniciu et al., 2000), then $d_{RS}(\boldsymbol{a}, \boldsymbol{b}) = \sqrt{1 - C(\boldsymbol{a}, \boldsymbol{b})^2}$. The Hellinger distance is defined as $d_H(\boldsymbol{a}, \boldsymbol{b}) = \sqrt{1 - C(\boldsymbol{a}, \boldsymbol{b})}$ (Hazewinkel, 2013), while the Bhattacharyya distance is defined as $d_B(\boldsymbol{a}, \boldsymbol{b}) = -\ln C(\boldsymbol{a}, \boldsymbol{b})$ (Bhattachayya, 1943). The total variation distance $d_{TV}$ is half the $\ell_1$-norm; therefore $d_{TV}(\boldsymbol{a}, \boldsymbol{b}) \le \frac{1}{2}\sqrt{s}\, d_E(\boldsymbol{a}, \boldsymbol{b})$. Combining these observations with (60) establishes (5):

$$\begin{cases} d_H = \sqrt{1 - \sqrt{1 - d_{RS}^2}}\,; \\ d_B = -\frac{1}{2}\ln(1 - d_{RS}^2)\,; \\ d_{TV} \le \sqrt{2s}\, d_{RS}. \end{cases}$$